PREDICTING DRIVER DISTRACTION: AN ANALYSIS OF MACHINE LEARNING ALGORITHMS AND INPUT MEASURES

An Undergraduate Research Scholars Thesis

by

TYLER WIENER

Submitted to the Undergraduate Research Scholars program at
Texas A&M University
in partial fulfillment of the requirements for the designation as an

UNDERGRADUATE RESEARCH SCHOLAR

Approved by Research Advisor: Dr. Tony McDonald

May 2018

Major: Industrial and Systems Engineering

TABLE OF CONTENTS

ABSTRACT

Predicting Driver Distraction: An Analysis of Machine Learning Algorithms and Input Measures

Tyler Wiener
Department of Industrial and Systems Engineering
Texas A&M University

Research Advisor: Dr. Tony McDonald
Department of Industrial and Systems Engineering
Texas A&M University

The research area on the detection and classification of distracted driving is growing in importance as in-vehicle information systems such as navigation and entertainment displays, which introduce sources of distraction for drivers, become more common in vehicles. To mitigate the potential consequences of distracted driving it is necessary for such systems to provide a means of detecting driver distraction and then responding appropriately. This study uses a machine-learning approach to develop classification models that detect and differentiate both cognitive and sensorimotor distraction among drivers, which were induced via secondary tasks in a simulator study. The inputs to these models are combinations of driving performance measures (e.g. brake force, lane offset, speed, and steering angle) and driver physiological measures (e.g. breathing rate, heart rate, and perinasal electrodermal activity), and the outputs are predictions of driver distraction (e.g. cognitive distraction, sensorimotor distraction, or normal driving). Various combinations of driving performance and driver physiological measures, multiple types of machine-learning algorithms, and a systematic feature extraction and reduction method called TSFRESH were used to develop the classification models. Results showed that the physiological measures did not provide significant information for detecting and classifying

driver distraction. Furthermore, no significant differences were found between the different machine-learning algorithms. Analyses on feature importance also revealed that driving performance measures including steering angle, lane offset, and speed were the most important indicators of distracted driving, and that features characterizing the extreme values, the variance and fluctuation, and the non-linearity and complexity of time series input were more informative for classifying driver distraction than other features. Conclusions suggest that distraction detection models gain more information from driving performance measures than physiological measures and that using features that characterize specific aspects of time series input is useful for classifying driver distraction.

ACKNOWLEDGEMENTS

I would like to personally thank my research advisor, Dr. Tony McDonald, for his continuous guidance throughout the course of this research project. Thanks also go to both the faculty and staff of the Industrial and Systems Engineering department for making my time at Texas A&M University a great experience.

I appreciate the Texas A&M Undergraduate Research Scholars (URS) program for the opportunity to formally participate in research and for providing helpful resources along the way.

Finally, thank you to my parents for providing endless encouragement and support during my undergraduate studies.

NOMENCLATURE

TSFRESH	Time Series Feature Extraction based on Scalable Hypothesis tests
RF	Random Forest
DT	Decision Tree
NB	Naïve Bayes
kNN	k-Nearest Neighbor
svmLin	Support Vector Machine (linear kernel)
svmRad	Support Vector Machine (radial kernel)
NN	Neural Network
phys	Driver physiology measures
db	Driving performance measures
dbPhys	Driving performance and driver physiology measures
dbBreath	Driving performance and breathing rate measures
dbHeart	Driving performance and heart rate measures
dbPerin	Driving performance and perinasal perspiration measures

CHAPTER I

INTRODUCTION

Driver distraction is a causal factor in many vehicle crashes. With the growing popularity for technologies such as in-vehicle information systems (IVISs), the problem of driver distraction is becoming progressively more significant. Numerous approaches including legislation, law enforcement, awareness programs, and vehicle monitoring have been employed to mitigate this global issue, however they have shown limited results in reducing the prevalence of distracted driving due to consistent injury and fatality rates (NHTSA, 2017). Other techniques that take advantage of distraction detection algorithms are promising, and could reduce distraction crashes by providing real-time interventions to drivers. The effectiveness of these technologies will be determined by their reliability and accuracy in identifying distraction; therefore it is important to thoroughly investigate all dimensions of algorithm development.

Consequences of Distracted Driving

The National Highway Traffic Safety Administration (NHTSA) reported 885,000 distraction-affected crashes in 2015 (14% of all crashes), resulting in 391,000 injuries and 3,477 deaths (NHTSA, 2017). Of the 3,477 reported fatalities, occupants of the vehicle comprised 84% while non-occupants comprised 16%. These crashes were identified as "distraction-affected" via a coding mechanism used in surveys based on police accident reports (PARs; NHTSA, 2017). A 2016 analysis on the Second Strategic Highway Research Program (SHRP 2) Naturalistic Driving Study (NDS) found that 68.3% of the 905 crashes reviewed involved some type of observable distraction. Such distractions included the use of in-vehicle and handheld devices, active interaction with passengers, and external distractions (Dingus et al., 2016). The SHRP 2

NDS involved the use of multiple onboard video cameras and sensors to gather continuous observations of over 35 million miles of driving data (Dingus et al., 2016).

One reason for the variation between the NHTSA and SHRP 2 NDS studies is that the NHTSA data was obtained from on-site and post-crash reports, which gather information only after a crash has occurred. Distraction does not always leave a physical trace, reporting systems vary across jurisdictions, and drivers are reluctant in admitting to distracted driving, therefore there may be considerable underreporting of distraction-affected crashes present in the NHTSA study. In contrast, the SHRP 2 NDS provided direct video observation of pre-crash and crash events as they occurred, which likely resulted in a more accurate estimate of the prevalence of driver distraction. In either case, distracted driving is an issue that affects many drivers and has the potential for serious consequences such as injury or death.

Global Impact

Distracted driving is also a global issue, with far-reaching impacts across people of different ages and genders. Among drivers aged 18-64 years, the prevalence of talking on a cell phone while driving at least once in the past 30 days ranged from 69% in the US, 21% in the United Kingdom, and 59% in Portugal, while percentages in the Netherlands, Belgium, Spain, France, and Germany ranged from 40% to 50% (Naumann, 2013). NHTSA reported that 36% of all distracted drivers involved in fatal crashes were in the age group 15-29, with lower percentages for older age groups (NHTSA, 2017). In the US few differences were found between sexes, as the average prevalence of using a cell phone in the past 30 days was consistent across males and females (~69% for talking and ~31% for texting or e-mailing) (Naumann, 2013). Variations across age groups and nationalities may be due to differences in the popularity of

devices like cell phones between age groups as well as differences in law enforcement and societal views between countries.

Definition of Driver Distraction

Driver distraction, as defined by Regan et al. (2009), involves a diversion of attention away from activities critical for safe driving toward a competing activity. Such examples might include talking, texting, or browsing on a cell phone, thinking analytically about the best route to take or how to solve a particular problem, or even undergoing emotional stress regarding loved ones or past experiences.

People engage in distracted driving for a number of reasons, some of which are inevitable and some of which are intentional. Inevitable sources of distraction involve situations that are out of the driver's control, such as an insect flying into the vehicle, an animal crossing a roadway, or perhaps even an emotional stress event. Intentional sources of distraction involve situations where the driver consciously decides to engage in a task that is not immediately important for safe driving. Such tasks can include texting on a cell phone, interacting with an in-vehicle information system (IVIS), or delegating cognitive resources to some specific problem. Drivers engage in inevitable sources of distraction due to circumstance, and often there is an element of interruption involved that attracts the attention of drivers, thus diverting it away from the driving task. On the other hand, drivers engage in intentional sources of distraction due to a voluntary decision, most often related to convenience or entertainment purposes. Inevitable and intentional distractions may vary depending on context. For example, sometimes the driver is compelled to direct their attention away from the immediate driving task in order to remain environmentally aware, such as reading road signs and locating other cars or pedestrians (Regan et al., 2009).

In addition to classifying driver distraction based on intentionality, distraction may also be classified on other characteristics, such as the source or means of distraction. Distraction can occur from sources such as objects, people, events, or activities, and can occur either inside or outside the vehicle. Also, distraction can happen through different means, such as disrupting control or diverting attention, which in turn can produce different outcomes, such as delayed response, reduced situation awareness, or increased risk (Regan et al., 2009).

Existing Mitigation Techniques

Several approaches exist to counteract and mitigate the impact of driver distraction including legislation and law enforcement, communications and outreach, and vehicle monitoring through the use of "telematics" devices. The United States, along with other countries, now enforce laws against texting, talking on a cell phone, and other sources of distraction for drivers. In addition to general legislation, many regions/states implement graduated driver licensing provisions or other specific restrictions that aim to reduce distracted driving by limiting the number of passengers and restricting cell phone use for young and/or novice drivers (Goodwin et al., 2013). Distracted driving enforcement models known as high visibility enforcement (HVE) have been used to deter cell phone use and distracted driving by increasing the perceived risk of arrest. Such models require dedicated law enforcement to actively pursue distracted drivers as well as paid and earned media support for the enforcement activity. While techniques such as graduated driver licensing provisions and high visibility enforcement have been effective, they require many resources and can involve a significant time investment to implement (Goodwin et al., 2013). General legislation, on the other hand, can be less costly but has been shown to be less effective in practice (Goodwin et al., 2013).

Employer programs as well as campaigns dedicated to medical conditions and medications also exist to increase the public's knowledge of distracted driving and its impact. Communications and outreach approaches to mitigating distracted driving have been shown to be rather ineffective at reducing the prevalence of distracted driving. This may be because although drivers know that they should be alert, in many cases either the distraction is out of the driver's control or it is deemed necessary and useful by the driver, in which case the voluntary driver behavior is hard to change (Goodwin et al., 2013). Furthermore, for campaigns to truly be effective they must be well developed, tested, and be long-term, all of which require substantial funding and support (Goodwin et al., 2013).

One recent approach to encouraging safe driving is the use of "telematics" devices by insurance companies. Such devices are used to understand the driving habits of drivers in order to identify and reward safe, and thus low-risk, driving behavior. Telematics devices are systems installed in vehicles that record information about driving habits, such as mileage, speed, acceleration/deceleration rates, time of day, etc. (Lowrey et al., 2011). These devices are typically installed via the onboard diagnostic port (OBD-II) that is present in most vehicles. Once installed, the telematics device collects information from the vehicle and then transmits that information to insurance providers using wireless phone networks (Lowrey et al., 2011). If driving behaviors such as hard stops, nighttime traveling, moving at high speeds, and averaging many miles per day are avoided, then drivers can qualify for discounts or reduced premiums from their insurance companies. In addition to potential savings on insurance costs, drivers can also become more self-aware of their driving behavior because telematics devices often offer the option to view the recorded data online. By reviewing this data, drivers can gain detailed insight on how they can become safer drivers. Programs such as Drivewise from Allstate and Snapshot

from Progressive are two examples of the use of telematics devices to encourage safe driving among drivers. A study on the acceptance of real-time and post-drive distraction mitigation systems found that perceived ease of use and perceived usefulness were primary determinants in drivers' intention to use distraction mitigation systems (Roberts et al., 2012). Although the study found that real-time systems were more obtrusive and less easy to use than post-drive systems (e.g. telematics devices), results did suggest that drivers found both systems useful (Roberts et al., 2012). Thus, if real-time distraction mitigation systems can be improved, there is evidence that such systems could be useful for mitigating distracted driving.

Overall there are several ways in which distracted driving can be combated, including legislation, law enforcement, communications and outreach, and vehicle monitoring. While some of these methods are more effective than others, ultimately the injury and fatality rates due to distracted driving has remained stable in recent years, suggesting limited impact and providing evidence that a new approach is necessary.

Distraction Mitigation Technology

A relatively new method aimed at mitigating and even preventing the consequences associated with distracted driving is to utilize technology that detects driver distraction and then adjusts vehicle systems or generates real-time feedback to the driver. This could be realized through an adaptive IVIS that constantly receives input from a detection algorithm. The task of distraction detection can be seen as a supervised machine-learning task in which training data labeled with instances of distraction is used to develop an algorithm that can predict future unlabeled data. Many such algorithms exist that take as inputs driver physiological information and/or driver performance measures and then output predictions on whether or not the driver is distracted. Driver physiological information includes heart rate, breathing rate, skin conductivity,

and eye gaze, all of which may be collected from instruments such as smart watches, medical sensors, or advanced camera systems. Driver performance measures include acceleration, brake force, steering, and lane position signals, which may be obtained from vehicle systems. Various combinations of input data may be used, including only driver performance, only driver physiology, or a mixture of both.

Detection algorithms typically use machine-learning techniques such as Bayesian Networks (Liang et al., 2018), Support Vector Machines (Liang et al., 2007), Random Forests (McDonald et al., 2013), k-Nearest Neighbors (Sathyanarayana et al., 2008), Neural Networks (Son & Park, 2016), and others to take the input data and try to characterize it into patterns that accurately differentiate each of the possible classes (e.g. distracted or not distracted). This characterization can be accomplished in numerous ways. Bayesian networks use graphical models to represent probabilistic relationships within the data (Heckerman, 1998). Support vector machines use complex functions to develop an optimal hyper-plane that separates the data (Burges, 1998). Random forests use a series of decision trees that split the data based on optimal predictors from randomly chosen subsets (Liaw & Wiener, 2002). The k-Nearest neighbor algorithm is based on minimum distances between data points (Larose, 2005). Finally, neural networks use layers of nodes with certain thresholds that determine whether or not data passes through to succeeding layers (Hammerstrom, 1993).

Often times several statistics, called features, about the data are calculated to help algorithms understand and characterize the patterns within the data. Such features include mean, standard deviation, frequency, duration, maximum, minimum, entropy, and other more complex features. The two main purposes of these features are to reduce noise in the raw data and to provide insight into the true underlying characteristics of the data. Raw data is often very noisy,

especially when that data deals with human behavior. Calculated statistics can combat such variations by providing a summarization of the data, focusing on the attributes of the data rather than simply the values. In this way a clearer picture of the patterns that exist within the data can be obtained, which significantly helps detection algorithms recognize and distinguish between each of the possible classes.

A detailed review of the literature on detecting driver distraction using various input data, machine learning algorithms, and features is discussed in Chapter II. Some of the limitations that exist in current detection algorithms, however, are that many algorithms consider only binary classification and do not use a systematic feature extraction and reduction approach when training. A binary classification task is one in which there are only two possible classes. An example of this would be an algorithm that detects whether a driver is distracted or not distracted. This information is useful, however it may be more helpful to determine the specific type of distraction that the driver is experiencing, such as cognitive distraction, sensorimotor distraction, or even emotional distraction. Specifying the type of distraction can allow adaptive IVISs to customize feedback to the driver that is well suited for the particular type of distraction, thus likely increasing effectiveness in mitigating that distraction. For example, an audible alert may be sufficient for a driver engaged in sensorimotor distraction, such as texting, but for a driver engaged in cognitive distraction that is lost in thought, perhaps a type of haptic feedback alert would be more effective in re-establishing the driver's attention. Also, using a systematic feature extraction and reduction approach allows algorithms to explore an expansive set of possible features and then optimize the set by reducing the number of features to a desired level. The reduction is based on feature importance and the overall performance of the algorithm. A systematic feature approach can also provide insight into what patterns the machine-learning

algorithm actually picks up on by identifying the features that have the greatest effect on model performance.

Research Objective

This study aims to achieve three primary goals. The first goal is to develop an algorithm that can accurately differentiate and classify cognitive distraction, sensorimotor distraction, and normal driving. This goal is aimed at going beyond traditional binary classification in order to distinguish between specific types of distraction. Identifying the specific type of distraction allows customized interventions or feedback to be provided by an adaptive IVIS system that is tailored to re-gain the driver's attention in the most effective way for that type of distraction.

The second goal is to test whether there are differences in model performance regarding the classification task when using different input data sets or different machine-learning algorithms. Using different input data sets and machine-learning algorithms helps to understand the importance of measures such as driver physiology and driver performance as well as the variations between machine-learning algorithms in the task of classifying driver distraction. This goal is aimed at highlighting the differences between input data types and machine-learning algorithms and ultimately identifying the most useful ones to use when classifying driver distraction.

The third goal of the study is to use a systematic feature extraction and reduction approach to explore numerous types of features and identify critical features for classifying driver distraction based on feature importance and overall model performance. This goal is aimed at providing a systematic analysis of the features that are used to characterize time series data and how and why certain features are more useful than others in the task of classifying driver distraction.

The remainder of this thesis is divided into five chapters that review the current literature on driver distraction detection, outline the methods used to conduct this study, list the results, provide discussion to put those results into context, and finally conclude with the most important findings from the study. In summary, data collected from a simulator study performed by the Texas A&M Transportation Institute (TTI) (Wunderlich et al., 2017) is used to systematically calculate numerous features and then train multiple algorithms on the classification task. Each algorithm is trained using only driver physiological measures, only driver performance measures, and several combinations of the two. The results of these algorithms are then analyzed to provide an understanding of the differences between various input data types and machine-learning algorithms as well as critical features for classifying driver distraction.

CHAPTER II

LITERATURE REVIEW

Many different approaches have been investigated in using machine-learning techniques to develop models that detect and classify driver distraction. A brief review of these approaches is provided in this section, particularly focusing on variations in input data types, algorithms, and features as they relate to the classification task.

Input Data Types

Input data types refer to the specific measures that are used as input to train distraction detection models. Common input data types are driving performance measures (Torkkola et al., 2004; Liang et al., 2007; Jin et al., 2012; Liang & Lee, 2014; Son & Park, 2016; Liang et al., 2018), face and eye tracking (Zhang et al., 2004; Liang et al., 2007; Sathyanarayana et al., 2008; Miyaji et al., 2009; Liang & Lee, 2014; Li et al., 2013; Ragab et al., 2014; Liu et al., 2016; Masood et al., 2018; Liang et al., 2018), and in some cases driver physiological measures (Sathyanarayana et al., 2008; Miyaji et al., 2009). Driving performance (e.g. brake force, lane position, speed, steering angle, etc.) measures are primarily used to gain insight into how distraction and workload levels affect drivers' control of the vehicle. Torkolla et al. (2004) used driving performance measures in combination with a multiple adaptive regression tree approach and, when compared to a state-of-the-art eye and head tracker, achieved a detection accuracy of 80%. It was concluded that driving performance measures alone could be used to detect driver inattention, thus avoiding the cost and complexity of adding driver monitoring sensors such as eye and head trackers. Jin et al. (2012) used velocity, acceleration, steering angle, throttle position, yaw angle, and angular velocity to detect driver cognitive distraction. Results from the

study suggested that compared with driver physical measures, using driving performance measures to detect distraction was more effective, simple, and of real time, suggesting that driving performance measures alone could be used to accurately detect driver cognitive state. Face and eye tracking (e.g. face orientation, eye gaze location, pupil diameter, eye closure, etc.) have been used to determine when the driver is not engaged in the critical tasks necessary for safe driving. Liang et al. (2007) used both eye movements and driving performance as inputs to develop a real-time classification model for detecting driver cognitive distraction. Results of the study showed that eye movements were clearly important as an input measure when comparing models trained using only driving performance and models trained with both driving performance and eye movements, ultimately recommending that both eye movement data and driving performance be included in as inputs to a distraction detection model. Driver physiological measures (e.g. breathing rate, heart rate, skin conductance, body motion, brain activity, etc.) have been tested to see if any additional discriminating information can be provided that improves the performance of distraction detection models. Miyaji et al. (2009) used physiological measures in addition to head movement and eye tracking to detect driver cognitive distraction. In the study, measures such as pupil diameter and heart R-waves from an electrocardiogram were shown to enhance the performance for distraction detection by adding the average values for the measures as pattern recognition features.

The vast majority of studies in the literature have used either driving performance, face and eye tracking, or a combination of the two as inputs to distraction detection models. There is room for further investigation of the effect of using driver physiological measures when developing distraction detection models, which is one of the goals of this study. Furthermore, although there is evidence to suggest that driving performance measures are more useful, that

face and eye tracking is more useful, and even that physiological measures are helpful in detecting and classifying driver distraction, most studies indicate that face and eye tracking are the most useful features for detecting driver distraction, followed by driving performance measures. Face and eye tracking data was not considered in this study so as to focus more on investigating the physiological measures that characterize a driver's biological response to distraction and the driving environment rather than their physical response.

Machine-learning Algorithms

Machine-learning algorithms are the techniques by which distraction detection models are trained and developed. Common algorithms used in the literature are Bayesian networks (Liang & Lee, 2014; Liang et al., 2018), decision trees (Torkkola et al., 2004; Zhang et al., 2004), Adaboost (Miyaji et al., 2009; Ragab et al., 2014), k-nearest neighbors (Sathyanarayana et al., 2008; Li et al., 2013), support vector machines (Ragab et al., 2014; Liang et al., 2007; Miyaji et al., 2009; Ersal et al., 2010; Jin et al., 2012; Liang & Lee, 2014; Li et al., 2013; Liu et al., 2016), and neural networks (Ragab et al., 2014; Ersal et al., 2010; Son & Park, 2016; Masood et al., 2018). Ragab et al. (2014) used a variety of machine-learning algorithms including Adaboost, hidden Markov models, random forest, support vector machine, conditional random field, and neural networks to develop a visual-based driver distraction recognition and detection model. An interesting finding of the study was that the experimental results showed superiority of the random forest classifier compared to the other classifiers. Liang and Lee (2014) conducted a study on the differences between a hybrid Bayesian network and an original dynamic Bayesian network (DBN). The hybrid network incorporated a layered algorithm with supervised clustering and, although there were no significant differences in classification accuracy or response bias, it was found that the layered algorithm significantly improved the computational efficiency from

the original algorithm in detecting driver distraction. Masood et al. (2018) used a convolutional neural network, along with images of in-vehicle environments, to develop a multi-class classification model of driver distraction. Ultimately the model was able to not only detect the presence of distraction but also identify the source of that distraction. Li et al. (2013) used both a kNN and SVM algorithm to develop binary and multi-class classification models. In addition to showing that each of the models was able to distinguish between normal and task conditions, a Gaussian mixture model (GMM) with promising results was developed to quantify the deviations of the driver behavior from expected normal patterns.

Although many different machine-learning algorithms and techniques are used for developing distraction detection models, the prevalence of support vector machines and neural networks largely outweighs other types of algorithms. That said, there is evidence to suggest that Bayesian networks and random forests show promising results in terms of computational efficiency and classification accuracy, respectively. Both a Bayesian network and a random forest algorithm, among a number of others, were included in this study to further evaluate any differences in model performance.

Features

Features refer to the calculated statistics or characteristics describing collected data that are fed as input to machine-learning algorithms in order to provide greater information. Studies in the literature typically use basic features such as mean, standard deviation, maximum, minimum, range, interquartile range, skewness, kurtosis, or other standard statistical measures (Torkkola et al., 2004; Zhang et al., 2004; Liang et al., 2007; Jin et al., 2012; Li et al., 2013; Liu et al., 2016), although sometimes more complex features are used such as eye movement temporal and spatial measures (Liang & Lee, 2014), eye closure and facial orientation (Ragab et

al., 2014), entropy and stationarity (Torkolla et al., 2004), and steering wheel reversal rate (Son & Park, 2016). Furthermore, feature reduction techniques, which are used to select optimal subsets of features for model training, are not very common in the literature. There are some studies that use different types of reduction techniques, however these studies are in the minority. Liang and Lee (2014), in addition to basic features for driving performance, calculated eye movement temporal and spatial features, including smooth pursuit movements, duration, direction, and speed, blink frequency, and vertical/horizontal fixation locations. A supervised clustering technique was also used in the study to group and reduce the feature set. It was shown that the supervised clustering technique was able to effectively integrate 19 distraction indicators into three feature behaviors. Furthermore, results suggested that the more complex features such as temporal characteristics of eye movement were the most predictive indicators of cognitive distraction, followed by spatial characteristics of eye movements and then the less complex features used for driving performance. Sathyanarayana et al. (2008) used principal components analysis (PCA) and linear discriminant analysis (LDA) for feature conditioning and reduction and found that these methods were useful for reducing feature sets to include only the most dominant features. Torkolla et al. (2004) incorporated entropy of error and multivariate stationarity features in addition to simpler running averages, differences, and variances in order to detect driver inattention in the absence of driver monitoring sensors and ultimately achieved a model with comparable performance to a state-of-the-art eye and head tracker. Ersal et al. (2010) used raw driving performance measures and vehicle dynamics to train a neural network and SVM classifier on classifying driver distraction under secondary tasks, using a resampling and filtering technique to reduce noise and prevent aliasing in the raw data input. Low classification accuracies resulted, however it was shown that different drivers were affected to different

degrees by the secondary tasks, suggesting that it may not be justified to classify all instances of driving with secondary tasks as distracted due to differences between drivers.

Many studies in the literature focused on developing driver distraction detection models use basic feature sets without feature reduction techniques designed to select optimal subsets of features. The studies discussed here are some of the exceptions, and were used to highlight the importance that complex features and systematic feature reduction techniques can have in developing distraction detection models. It is one of the goals of this study to expand on such investigations by using a relatively new feature extraction and reduction technique, called TSFRESH, to analyze any effects on model performance and to identify critical features for detecting and classifying driver distraction.

Overall there are many different approaches and techniques used in the literature for detecting and classifying driver distraction. Some of the common conclusions that the literature offers are that eye and face tracking are important and informative input measures for classification models, that support vector machines and neural networks are proper algorithms for distraction detection, and that standard statistical measures serve as good features for identifying driver distraction. Opportunities for further expansion and investigation include the effects of driver physiological measures (not related to face and eye tracking), differences between other machine-learning algorithms (including random forest and Bayesian networks), and potential insights from using more complex feature extraction and reduction techniques. All of these opportunities have been identified as goals of this study.

The remainder of this thesis describes the simulator data used in the study, details the methods used to train various machine-learning algorithms, compares the results of each of the

models, and finally expands on those results through a discussion focused on real-world driving context and impact for future research.

CHAPTER III

METHODS

This section reviews the simulator study, the methods of inducing distraction, the resulting dataset, and the procedure by which the dataset was used to calculate features and train machine-learning algorithms in this research. The simulator study was conducted on a Realtime Technologies Inc. (RTI) driving simulator (Wunderlich et al., 2017) at the Texas A&M Transportation Institute (TTI). Distracted driving was induced with cognitive, emotional, or sensorimotor secondary tasks. The data collected from the study included driver physiological measures, driving performance measures, and survey responses. The resulting dataset was pre-processed for model training using the statistical computing environment R and features were calculated using a package in Python called TSFRESH (Time Series Feature Extraction based on Scalable Hypothesis tests; Christ et al., 2017). The machine-learning algorithms were trained using the caret package in R (Kuhn, 2017) and then subsequent testing and analysis was also performed using R.

Simulator

Figure 1 shows the RTI driving simulator used to measure driver performance under varying degrees of distraction. Included in the simulator was a vehicle seat, steering wheel, accelerator and brake pedals, three screens for displaying the driving environment, and a speaker system to provide ambient roadway noise. Drivers' responses to roadway situations were measured by the simulator, including steering wheel position, accelerator and brake pedal position, velocity, time to lane crossing, time to an upstream vehicle, and lane position. Driving data was collected at 60 Hz and then later aggregated to 1 Hz for analysis.

Figure 1. TTI Driving Simulator (Wunderlich et al., 2017)

Study Procedure

Participants were recruited through email and flyer solicitations from the Bryan and College Station, Texas communities. Admission was restricted to individuals having a valid driving license, normal or corrected to normal vision, at least one and a half years of driving experience, and no medications that could affect driving ability. Two age groups: 18-27 (younger) and 60 and above (older) were represented in the study.

A total of 78 participants volunteered for the study. One individual quit due to motion sickness and data for 9 others were not recorded properly due to technical issues, resulting in a total of 68 participants in the simulator study. For each participant and every physiological and performance measure, the experimenters determined the state of the collected data as being either valid, missing due to technical reasons, invalid due to noise, not present due to experiment design, redacted due to Institutional Review Board (IRB) restrictions, or missing due to the presence of facial hair. Analyses could not be performed on 9 male participants because the presence of facial hair caused the perinasal perspiration signal to be problematic. For the

purposes of this research 11 other participants were removed due to missing or invalid data. Thus, a total of 48 participants (21 male, 27 female) were considered in this study. Individuals were balanced across age, with 10 males/14 females in the younger group and 11 males/13 females in the older group. The emotional distraction portion of the simulator study was also removed from the current analysis due to limited confidence from the experimenters in the effectiveness of the method used to induce emotional distraction.

The simulator study was a controlled experiment including eight experimental sessions. The participants were randomly assigned to two groups: Nonloaded and Loaded. These groups only affected the last session in the experiment. All other experimental sessions were the same for both groups. Three questionnaires (Biographic, Trait Anxiety, and Personality Type) were completed prior to the eight experimental sessions to understand different characteristics of the participants that might affect their driving performance. The experimental sessions were conducted in the simulator, where participants drove through various "driving worlds" that were created by the experimenters. More details on the creation and programming of the driving worlds can be found in Wunderlich et al. (2017). Between each of the experimental sessions was a two-minute break during which participants completed the NASA Task Load Index (TLX) for the preceding session.

The first experimental session was a baseline session in which individuals sat in a dimly lit room, quietly listening to soothing music for five minutes. The next session was the practice session, which allowed participants to familiarize themselves with the simulator throughout an 8 km drive on a straight four-lane highway with varying speed limits and traffic density. After the practice session was the relaxing drive, which consisted of a 10.9 km straight section of a four-lane highway with a speed limit of 70 kph and light oncoming traffic (3 vehicles/km). A forced

lane change was also introduced approximately 5.2 km into the drive. The next four drives were loaded drives. These drives all featured the same challenging driving conditions (construction zones, heavy oncoming traffic, buildings, and a forced lane change), however the order of the drives for each participant was randomized. One of the loaded drives presented no additional stressor while the other three presented an additional stressor in the form of a secondary task (cognitive, emotional, or sensorimotor). The secondary task was forced in two non-consecutive phases of the drive, which were implemented first at 1.2 km into the drive and then at 7.2 km into the drive, each lasting for approximately 3.2 km. All loaded drives were on the same 10.9 km section of a four-lane highway with a posted speed limit of 70 kph, oncoming traffic density of 12 vehicles/km, 2 buildings/km, and a forced lane change 5.2 km into the drive. The normal loaded drive (LD_0) consisted of only the construction activity and no additional stressor. The cognitive drive (LD_C) consisted of mathematical questions in one phase of the drive and analytical questions in the other phase, which were posed orally by the experimenter in a randomized phase order across participants. The emotional drive (LD_E) included emotionally stirring questions posed orally by the experimenter in two phases, which were randomly ordered across participants. One phase included less pointed questions while the other phase included more pointed questions. The sensorimotor drive (LD_M) had participants text back words that were sent one by one to the participant's smartphone, again in two phases. The last experimental session was another drive in the simulator, this time introducing an unintended acceleration event (UAE) to the participants. Individuals drove on a 3.2 km highway section identical to the last 3.2 km segment in each of the loaded drives. The Nonloaded group (randomly assiged at the beginning of the experiment) did not engage in any secondary activity during the drive, while the Loaded group drove under mixed stressors for the last 2 km of the drive. Towards the end of the

drive was an intersection with a red light. Prior to the light turning green, the participant's vehicle malfunctioned and propelled the vehicle forward, putting it on a collision course with another vehicle that had entered the intersection. The participants had five seconds to react before a collision occurred.

Data Collection

Data was collected continuously throughout each drive in the simulator. The data consisted of both driver physiological data as well as driving performance data.

The physiological data was measured with several biological instruments and included perinasal electrodermal activity (EDA), palm EDA, heart rate, breathing rate, and eye tracking signals. The instruments used were thermal imaging cameras and algorithms, a chest strap heart sensor, galvanic skin response (GSR) sensors, and an eye tracking system. The palm EDA signal was not considered in this study because the experimenters identified that many of the participants had either missing or invalid data for this signal. The eye tracking data was not considered because the focus of this study was to test if adding physiological measures that characterize the driver's biological response to the environment helps in the classification of driver distraction.

The driving performance data was measured from the simulator itself and included acceleration, brake force, distance, lane offset, lane position, speed, and steering signals. Acceleration was not included in this study because the participants were instructed to maintain speeds according to speed limits and traffic demands, thus this measure should not be related to distracted driving. Distance is simply an aggregate of the miles travelled, and also was removed because it should not be indicative of distracted driving. Finally, lane position is a different form of the lane offset measure, so it was removed due to redundancy.

All data was aggregated and compiled to a dataset freely available on the Open Science Framework (OSF). More details on this study can be found in Wunderlich et al. (2017). The following sections detail the procedure used to preprocess the data, extract and reduce feature sets, and train and test the various classification models. Figure 2 provides an outline of this procedure.

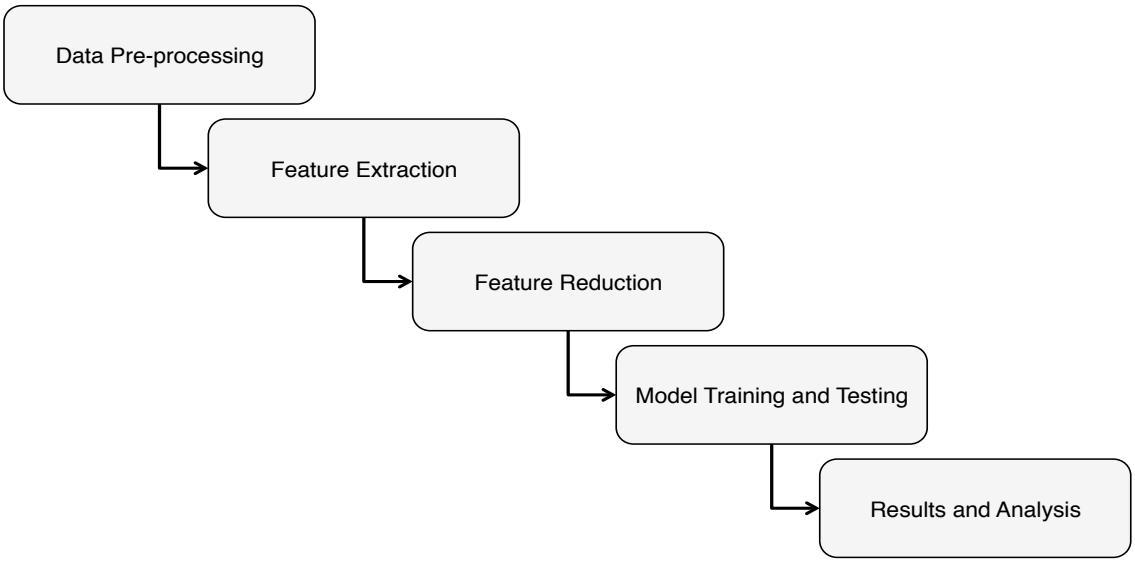

Figure 2. Model Training and Testing Procedure

Data Pre-processing

All data pre-processing was done in the statistical computing environment R. The dataset published on the Open Science Framework includes a comma-separated values (csv) file for each of the 68 participants in the simulator study. This csv file includes the raw values of the driver physiological and driving performance measures for a given participant during all drives in the study. All files were combined by vertical concatenation in order to have the data for all 68 participants in one file. The 9 male participants with facial hair and the 11 participants with either missing or invalid data were removed. The experiment identification variables, including

stimulus (the secondary task) and drive (the experimental session), were decoded from their numerical representations to text following the scheme described on the OSF wiki page for the experiment. All portions of the experiment other than the loaded drives were removed, as they were not important for training the models. The emotional distraction drive of the experiment was removed for all participants, however, due to questions of its validity by the experimenters.

A key indicating the participant and the drive was created to identify the different portions of the experiment. Instances in time that contained one or more NA values for the various driver physiological and driving performance measures were removed to ensure that only full, valid data was included in training the models. Each of the included physiological and performance measures were normalized using a z-score transformation to scale and center the data. This was done to remove any biases that stem from differences in units and ranges between the measures. Since some machine learning algorithms are sensitive to magnitude, such as Support Vector Machine and k-Nearest Neighbor, normalizing the data is a key step before training the models.

The next important aspect in pre-processing the data was to partition the data into equivalent segments of time. These time segments, called windows, serve as inputs to the detection algorithm for which a corresponding output, in the form of a prediction of driver state, is outputted. A real-time intervention system cannot realistically provide feedback based on instantaneous input, therefore a particular time window is needed over which a detection algorithm can evaluate the input and then produce its output. Non-overlapping, fixed-interval segments of 30 seconds were chosen as the partitioning windows for the data in this study. Although Liang et al. (2018) discussed that longer windows may undermine models' performance because there are fewer training instances and there may be a delay in detecting

distraction, Liang et al. (2007) provided some evidence that longer windows improved model performance, and so the initial window size of 30-seconds was maintained for this study. The windows were created chronologically in each of the drives for every participant in the dataset. Because the windows were created chronologically in each of the drives, there were some windows that spanned across the portions of a drive that transitioned between no secondary task and a secondary task or vice versa, from a secondary task to no secondary task. These windows were kept in the dataset, and were handled by labeling the window with either the majority class label or, in the case of a tie, the first class label in the window (details below). Incomplete windows from the ends of drives were removed from the dataset.

For training the models, the windows must also be labeled with the corresponding driver state in that window so that the algorithm can associate the patterns in the data to their corresponding class. The corresponding class for a given window was determined from the stimulus, or secondary task, that was present in that window. For example, if a participant was performing the sensorimotor distraction drive, and was engaged in the secondary task of texting on a cell phone over the course of a window, then the corresponding class for that window would be designated as sensorimotor distraction. Each 30-second window in the data contained either one or two classes, depending on whether or not that particular window crossed a transition from secondary task to no secondary task or vice versa. This situation was handled in one of two ways. If the two classes were uneven in terms of presence in the window, then the majority class in the window was selected as the overall class for that window. If the two classes were perfectly balanced, meaning both classes were present for exactly 15 seconds of the 30-second window, then the first class was selected as the overall class for that window. This method was chosen in order to account for either delays in or lingering effects of the distracting secondary tasks. In one

case it might take a participant several seconds to fully engage in the secondary task, and in the other case the effect of the secondary task might linger for a few seconds after it is completed. Further post-study analyses are needed to analyze any effects of these "transitional" windows on model performance, specifically whether or not misclassification of distraction was more prevalent for transitional windows than for non-transitional windows. After labeling the windows with their corresponding class, all of the windows associated with the non-loaded portions of the loaded drives were removed from the dataset. These windows refer to the portions of the loaded drives where no stimulus, or secondary task, was performed. Instead of labeling these windows as "normal driving," they were removed in order to avoid confusing the algorithms by labeling a potentially distracted scenario as normal due to a possible lingering behavioral effect from the secondary task.

Finally the dataset was split into training and testing sets and then checked for class imbalances. To select the training and testing sets, random numbers were generated for each participant in the study. This list was then sorted in ascending order and the first 5 participants were taken to be part of the testing set. Five participants were chosen in order to achieve a 90% to 10% split between the training and testing sets. The testing set was also checked for balance across gender and age groups to ensure a representative sample, with 1 participant from each of the Male Younger (MY), Male Older (MO), and Female Younger (FY) groups and 2 participants from the Female Older (FO) group. To check the training and testing sets for class imbalances, the counts of each class present in the dataset were generated. The dataset was down sampled to match the least frequently occurring class in the dataset, which was sensorimotor distraction. This was done primarily to remove bias in the algorithms for predicting one class over another due to differences in frequency. Balancing the datasets is also important, for the same reason,

when using prediction accuracy as a metric to optimize the algorithms. After down sampling in both the training and testing sets, there were 421 windows of each class (1263 total) in the training set and 49 windows of each class (147 total) in the testing set.

Feature Extraction and Reduction

The next step in the data cleaning and pre-processing phase was to extract features using the TSFRESH package for the popular programming language Python. TSFRESH stands for Time Series Feature Extraction based on Scalable Hypothesis tests. The package is used to automatically extract hundreds of features from time series data, which can be used in constructing statistical or machine-learning models for regression and classification tasks. TSFRESH features include a variety of distributional (e.g. mean), non-linear (e.g. approximate entropy), spectral, Fourier coefficients, wavelet, polynomial, and miscellaneous features (e.g. time reversal asymmetry) that span a wide range of complexity. One example of how these features relate to characterizing driving time series data in particular is the mean absolute change feature. This feature calculates the mean over all of the absolute differences between subsequent time series values and is similar to the steering reversal rate measure that has been associated with the evaluation of driving performance (Macdonald & Hoffman, 1980).

All that is required to extract features for a classification task in TSFRESH is a time series dataset containing the raw data values and a class dataset containing the class labels that correspond to those raw data values. The specific format of these input datasets is outlined in the TSFRESH documentation. The required format is a time series dataset with an ID column, a time column, and the columns containing the actual values. To achieve this format, unique indices were assigned to the training instances based on their key (composed of the participant and the drive). These indices served as the ID column for TSFRESH. The time column was simply the

time into window column that was calculated in the data pre-processing stage. The class dataset was created by simply forming a vector of the class labels for each of the unique windows in the training dataset. The order of the class labels must match the order of the windows in the training dataset so that TSFRESH knows which class labels go with which training windows. In addition to the training dataset and class dataset, another dataset for the time series data from the testing group is also needed in order to extract the same features as in the training set and eventually predict the classes for the training group.

Once these datasets are built, simply loading them into Python and running the appropriate function will output the list of extracted features. The two TSFRESH functions used for feature extraction in this study are called "extract_features" and "extract_relevant_features." The "extract_features" function simply calculates all of the hundreds of features that are offered in TSFRESH. This function was used to extract all of the features for the testing dataset, which was reduced to match the features in the final training set after feature reduction was performed in R. The "extract_relevant_features" function, on the other hand, not only calculates the hundreds of features that TSFRESH offers but it also performs feature filtering to remove irrelevant or uninformative features. This function was used to extract features from the training dataset that were relevant in regards to the classification task. Feature filtering in TSFRESH works by evaluating the explaining power and importance of each characteristic for the classification task at hand and uses a multiple test procedure from the theory of hypothesis testing. In particular, significance tests are performed for each feature, resulting in a vector of p-values that quantify the significance of each feature for predicting the target under investigation (i.e. the class dataset). The vector of p-values is then evaluated based on the Benjami-Yekutieli procedure in order to decide which features to keep. More information on the feature filtering

process used in TSFRESH can be found in Christ et al. (2016) and Benjamini & Yekutieli (2001).

After having calculated and extracted features from TSFRESH, further feature filtering and reduction was performed to decrease the amount of features in the final training dataset. There are two primary reasons for further reducing the feature set. First, as the ratio of width to height (i.e. number of features to number of training instances) in the training dataset grows, the likeliness that algorithms pick up on spurious correlations increases. The second reason for reducing the size of the feature set is to decrease computational complexity when training the algorithms. As the size of the input dataset increases, so does the time that it takes to train the classification models. Therefore it is preferable to retain a small feature set in order to obtain short training times, yet we must also be sure to include an adequate amount of information so that the algorithm can accurately distinguish between classes. To accomplish this second phase of feature filtering, several manipulations in R were performed. All of the features that contained one or more NA or infinite (Inf) values, which may be generated by TSFRESH depending on how the specific feature is calculated, were removed from the final feature set. Furthermore, all of the features that exhibited either zero or near zero variance in their values were removed. This was done by identifying the features that either had one unique value (zero variance) or had both of the following characteristics: very few unique values relative to the number of samples (less than or equal to 10%) and a large ratio of the frequency of the most common value to the frequency of the second most common value (greater or equal to 95/5). The reason for removing features with zero or near zero variance is that they are likely not very useful in distinguishing between classes for the classification task because they have a similar set of values for each the classes. Finally, all features that were highly correlated with another feature (greater than 0.9)

were removed. This process was done to remove any redundant information or spurious correlations in the feature set and was achieved by removing the feature with the largest mean absolute correlation in a pair of highly correlated features.

As a last step in the feature reduction phase, the feature sets for the training and testing sets were made to match each other for later prediction and testing using the classification models. To accomplish this, the feature set for the testing set was reduced to match the features in the final training set. Then, any features containing NA values in the testing set were removed from both the training and testing set, ultimately yielding a finalized feature set that was the same between training and testing datasets. The ID column was then reverted back to reflect its particular key (participant and drive) and window values. The feature datasets were now ready to serve as inputs to the machine-learning algorithms so that classification models could be trained.

Model Training and Testing

Now that the feature sets were finalized for both the training and testing sets, each of the various models could be trained using the different input datasets and machine-learning algorithms. Six different input datasets were used, which consisted of combinations of driver physiological measures including breathing rate, heart rate, and perinasal perspiration, and driving performance measures including brake force, lane offset, speed, and steering angle. The input dataset combinations were as follows: only driver physiological measures (phys), only driving performance measures (db), both driver physiological and driving performance measures (dbPhys), driving performance and breathing rate measures (dbBreath), driving performance and heart rate measures (dbHeart), and finally driving performance and perinasal perspiration measures (dbPerin). The purpose of having these specific combinations was to determine if physiological measures themselves were sufficient in describing driver state in terms of

distraction as well as to see what type of effect introducing different physiological measures to driving performance measures would have on model performance. Seven different machine-learning algorithms were used, which were chosen based on popular algorithms for both general classification tasks as well as specifically classifying driver distraction. The machine-learning algorithms were as follows: random forest (RF; Liaw & Wiener, 2002), decision tree (DT; Therneau et al., 2017), naïve Bayes (NB; Michal Majka, 2018), k-nearest neighbor (kNN; Kuhn, 2017), support vector machine with a linear kernel function (svmLin; Meyer et al., 2017), support vector machine with a radial kernel function (svmRad; Karatzoglou et al., 2004), and finally neural network (NN; Venables & Ripley, 2002). In addition to being chosen based on popularity, the algorithms were also chosen because there are certain advantages with each of them. For example, random forests incorporate an ensemble technique, decision trees are easy to understand and interpret, naïve Bayes classifiers are simple and typically converge quickly, k-nearest neighbor algorithms are straightforward and robust to outliers, support vector machines can model non-linear data, and neural networks are able to detect complex underlying relationships. Another reason for including several different machine-learning algorithms was to test if for significant differences in performance between the algorithms in terms of classifying driver distraction.

In total there were 42 models trained and tested on the classification task. For each model, the appropriate training and testing sets were constructed by extracting the necessary features from the finalized training and testing feature sets. This was done using R to pull out all features that began with the name of the measure that they were calculated for. For example, if the input dataset combination was driving performance and heart rate measures, then all of the features that were calculated for the brake force, lane offset, speed, steering, and heart rate

measures that were in the finalized feature sets would be used for the training and testing sets for that model. It was assumed in this process that the feature sets generated from TSFRESH for all of the measures at one time would be similar, if not equal, to those generated for the measures individually, which is based on the fact that TSFRESH uses a consistent algorithm for testing feature significance and removing irrelevant features. All models were trained using the functions provided in the caret package in R, which was built to provide a standardized interface for training machine-learning models. The caret functions simply require a training set with the values of the training data (i.e. the features), a class dataset (i.e. the class labels), and parameters to specify the machine-learning algorithm and tuning settings. More details on the caret package can be found on the package website. All models were trained using ten-fold cross validation, which is a technique used to randomly partition the training dataset into equal size subsamples and then hold one of the subsamples for validation and testing purposes. The reason for using ten-fold cross validation was to make up for having a relatively small dataset by using a resampling technique as well as to gain an understanding of model performance during training and then tune the model's parameters accordingly for optimization. All procedures performed in R that included the use of random numbers were seeded to ensure reproducibility in the results.

Once the models were trained, predictions were made on the data in the testing set using functions in R. The resulting predictions, along with various measures of performance such as the confusion matrix, average accuracy, mean area under the receiver operating characteristic (ROC) curve, and mean sensitivity and specificity, were compiled and saved for analysis. The next section lists the results of testing the various models along with the analyses that were performed.

CHAPTER IV

RESULTS

The trained models were evaluated by assessing their prediction of the testing data. This prediction gives an estimate of the generalizability of the algorithms to new, previously unseen data. Algorithms were evaluated based on their accuracy and Area Under the receiver operating characteristic Curve (AUC). One algorithm, support vector machine with the linear kernel (svmLin), was not included in the AUC analysis because AUC was not calculable with the library's algorithm. Overall results for the algorithms are summarized in Table 1 and 2.

Table 1. Model Accuracy

	Machine-learning Algorithm						
Input Data	RF	DT	NB	kNN	svmLin	svmRad	NN
phys	0.52	0.39	0.42	0.39	0.46	0.43	0.29
db	0.69	0.59	0.64	0.64	0.69	0.58	0.64
dbPhys	0.66	0.58	0.65	0.48	0.66	0.46	0.55
dbBreath	0.67	0.58	0.67	0.54	0.67	0.48	0.61
dbHeart	0.68	0.59	0.64	0.66	0.74	0.57	0.65
dbPerin	0.67	0.59	0.61	0.58	0.73	0.48	0.50

Table 2. Model AUC

	Machine-learning Algorithm					
Input Data	RF	DT	NB	kNN	svmRad	NN
phys	0.66	0.57	0.64	0.54	0.62	0.56
db	0.86	0.75	0.82	0.8	0.73	0.84
dbPhys	0.86	0.75	0.82	0.71	0.65	0.78
dbBreath	0.85	0.75	0.83	0.74	0.64	0.79
dbHeart	0.88	0.75	0.83	0.81	0.73	0.77
dbPerin	0.85	0.75	0.8	0.76	0.66	0.71

The remainder of the section is divided into three parts that describe the input data types, the machine-learning algorithms, and the features, respectively.

Input Data Types

The 7 different machine-learning algorithms were trained using each of the 6 input datasets. Figure 3 shows the average accuracy, along with 95% confidence intervals, for each of the models grouped by input data type. Figure 4 shows the mean AUC for each of the models grouped by input data type.

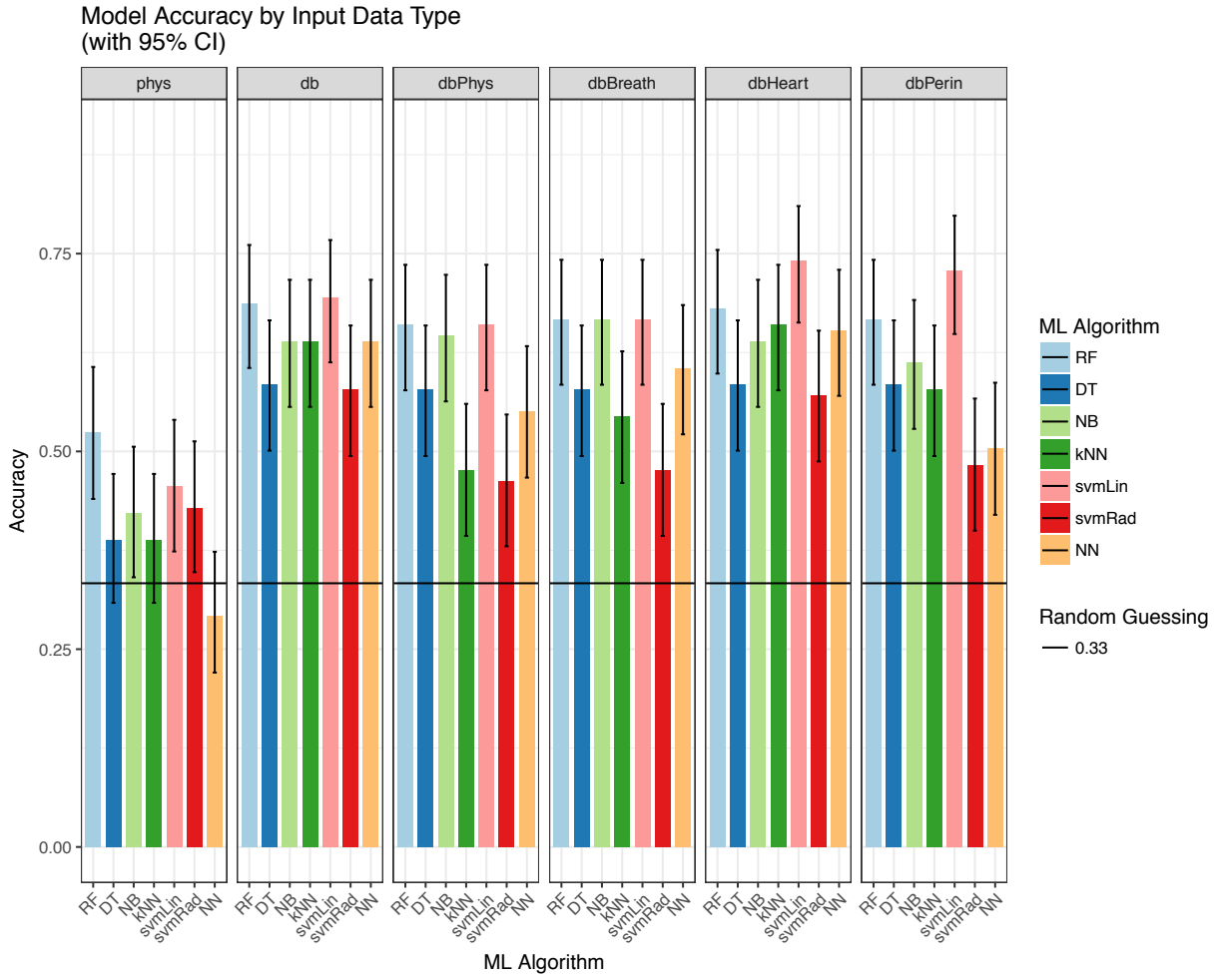

Figure 3. Model Accuracy by Input Data Type

The 95% confidence intervals in Figure 3 illustrate that all but three of the models performed statistically better than random guessing (33%; the horizontal black line in the figure). All three of the models that did not perform statistically better than random guessing were trained using only driver physiological measures (phys input data type). The model AUC values shown in Figure 4 illustrate similar trends. These figures indicate that physiological measures alone are not sufficient to accurately detect and classify driver distraction. Furthermore, there is no apparent difference in model accuracy achieved by the models that used driving performance only (db input data type) and the models that used additional physiological measures (dbPhys, dbBreath, dbHeart, dbPerin). This suggests two things. First, driving performance measures must dominate physiological measures in the task of classifying driver distraction since accuracies did not improve by adding physiological measures. Second, either the physiological measures provide very little insight into the driver's state of distraction or they all provide similar information, which resulted in approximately equal changes in model performance between the input data types.

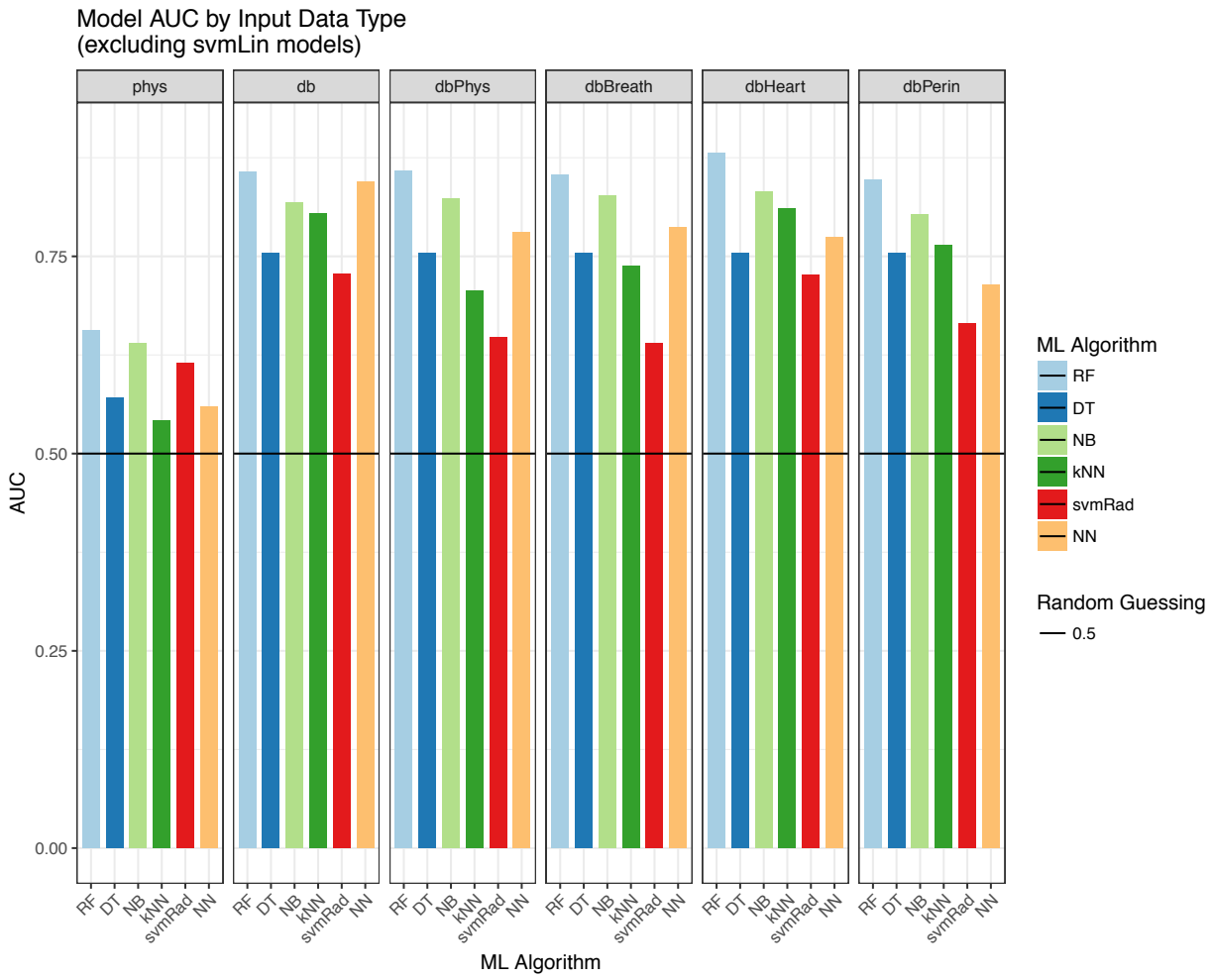

Figure 4. Model AUC by Input Data Type

Machine-learning Algorithms

Figure 5 shows the average accuracy, along with 95% confidence intervals, achieved by each of the models grouped by machine-learning algorithm. Figure 6 shows the mean area under the receiver operating characteristic curve for each of the models grouped by machine-learning algorithm.

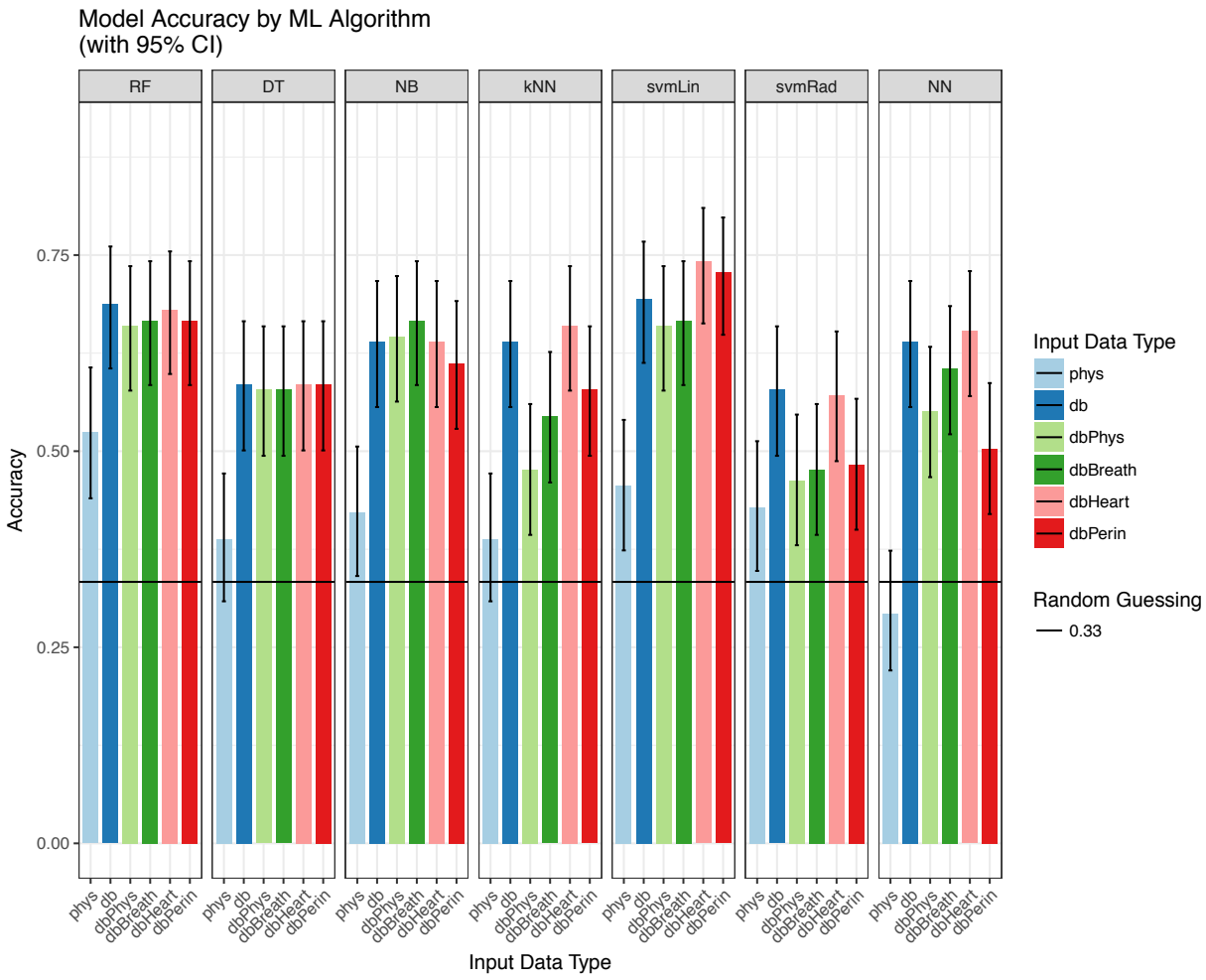

Figure 5. Model Accuracy by Machine-learning Algorithm

One interesting result highlighted in Figure 5 is that, excluding the phys models, there is no significant difference between the vast majority of machine-learning algorithms. Other than a few extreme cases, most of the confidence intervals for the model accuracies overlap with each other. It does appear, however, that certain algorithms are more consistent across different input data types than other algorithms. Random forest, although not the most accurate or the most consistent, is the one machine-learning algorithm that achieves high accuracy and high consistency. Random forest is also the only machine-learning algorithm in this set that

incorporates ensemble classification techniques. While ensemble techniques may not be the source of the random forest's performance levels, they may have something to do with the fact that the random forest models were both high performing and consistent. Another interesting result from looking at the performance between machine-learning algorithms is that generally speaking increasing model complexity does not achieve better performance. The more complex models in this set are random forest, support vector machine, and neural network whereas the simpler models are decision tree, naïve Bayes classifier, and k-nearest neighbor, however the performance levels of these models does not follow the hierarchy of model complexity. While this result could be due to the parameter tuning and model optimization techniques that the caret package uses in R, it is certainly an interesting outcome of testing the models on the driver distraction classification task. A similar picture is illustrated with model AUC in Figure 6.

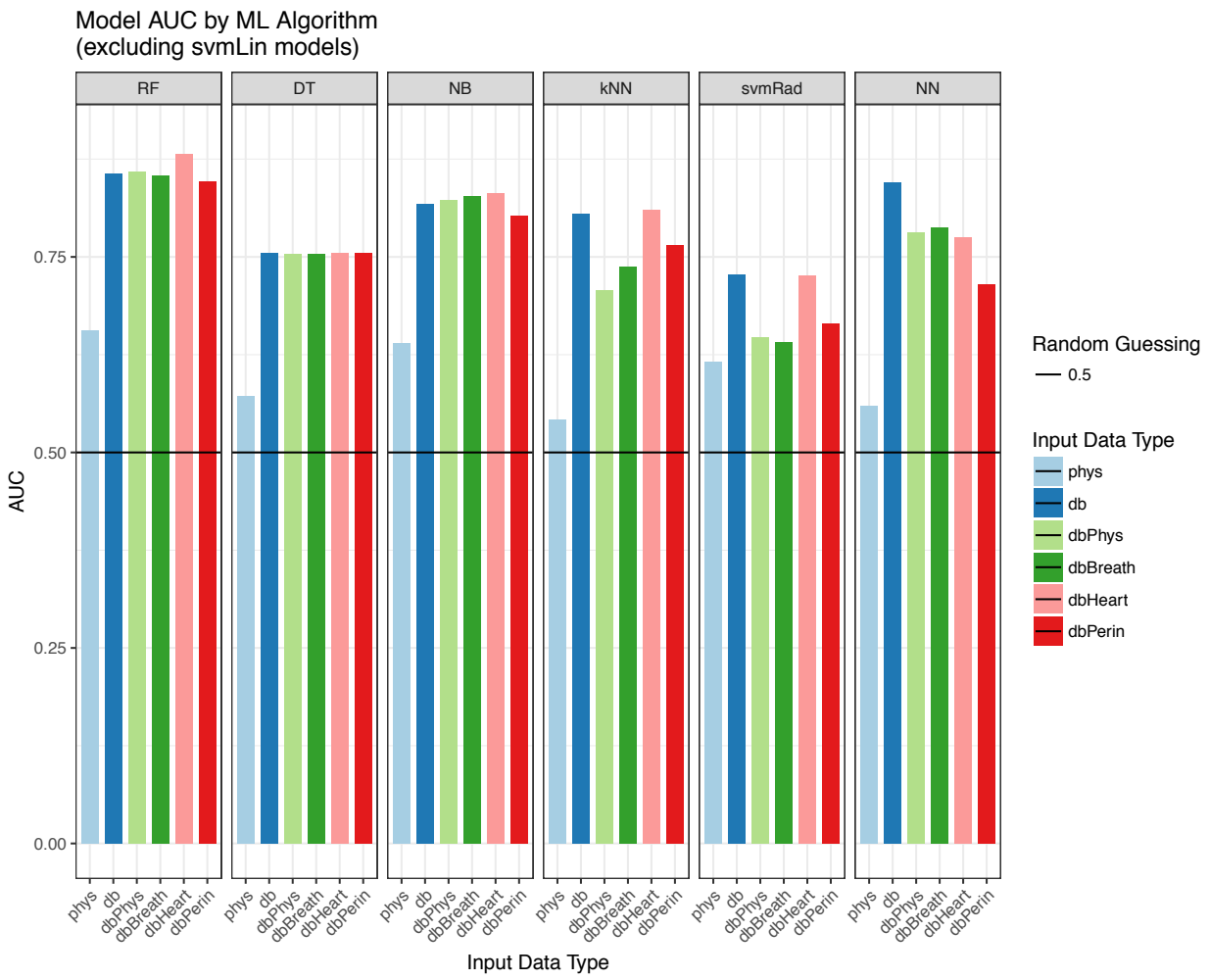

Figure 6. Model AUC by Machine-learning Algorithm

Features

Feature importance calculations were also compiled for each of the models, using the standard methods defined in R for each of the machine-learning algorithms. For both simplicity and brevity, the random forest models were selected as the focus of the remainder of this discussion. This is because the random forest models were both high performing and consistent. Furthermore, random forests provide some advantages in understanding feature importance because they are comprised of decision trees, a structure for which variable importance measures

are well defined. In this study, feature importance was calculated using mean decrease in accuracy, which is the average decrease in model accuracy, or performance, that results after removing a particular feature from the model's dataset (Liaw & Wiener, 2002). For example, a higher mean decrease in accuracy indicates a more important feature because the model performed far worse after excluding the feature from the dataset than it did when retaining the feature. Lastly, there were little to no significant differences between the machine-learning algorithms, therefore it is believed that any general conclusions drawn from the random forests would apply to other models as well. The phys input data type was excluded from this discussion because all models trained using only physiological measures performed far worse than the models trained with driving performance measures.

Figure 7 shows the top 10 features that were present in each of the random forest models for all of the input datasets that contained driving performance measures (i.e. db, dbPhys, dbBreath, dbHeart, dbPerin).

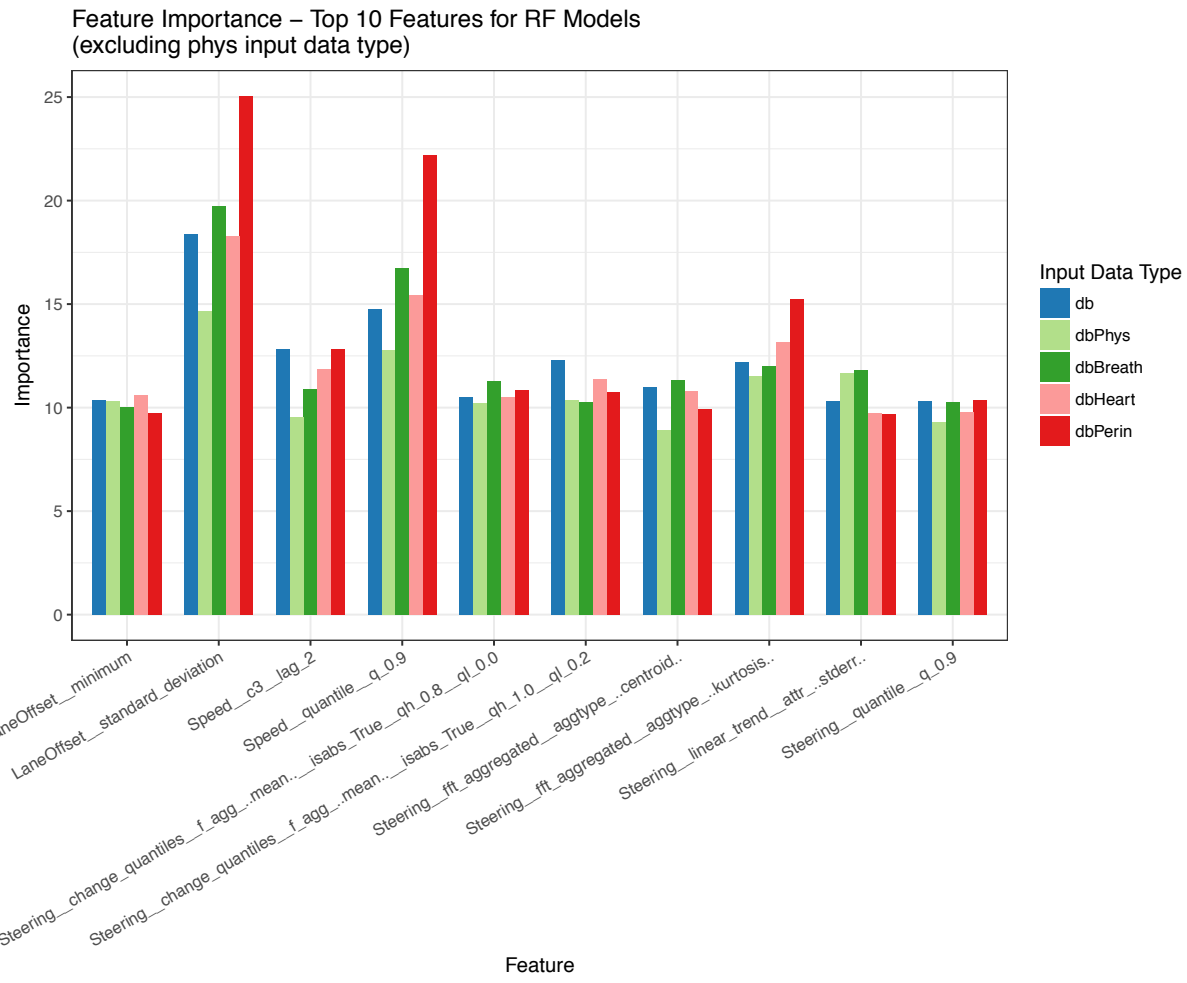

Figure 7. Top 10 Features Present in all RF Models (excluding phys input data type)

The intersection of the top features for each of the five considered input datasets was taken in order to identify the top 10 features present across all of them. As shown in the graph, the top features across these five input data types consisted of lane offset, speed, and steering angle features. This makes sense because the driving performance (db) model only contained features regarding driving performance. It is interesting, however, that the three specific measures of lane offset, speed, and steering were most prevalent. Both lane offset (i.e. lane position) and steering angle have been well defined in the literature and have been shown to relate to driver distraction.

Lane offset, or generally lane position, and steering angle tend to fluctuate under physical distractions, such as sensorimotor distraction in this study, while remaining more consistent and level under mental distractions, such as cognitive distraction in this study. Speed is another measure that has been studied in the literature on driver distraction. Speed has been linked to compensatory behavior of distracted drivers. In other words, distracted drivers have been shown to reduce their speed as a way to "offset" decreased control of the vehicle due to distraction. To gain a clearer understanding of which specific features were important for the individual models and why they were important, a deeper dive into the distributions of the top features for the models was performed.

Figure 8 shows the feature importance levels of the top 100 features for the random forest model trained with all of the driver physiological and driving performance measures (dbPhys input data type).

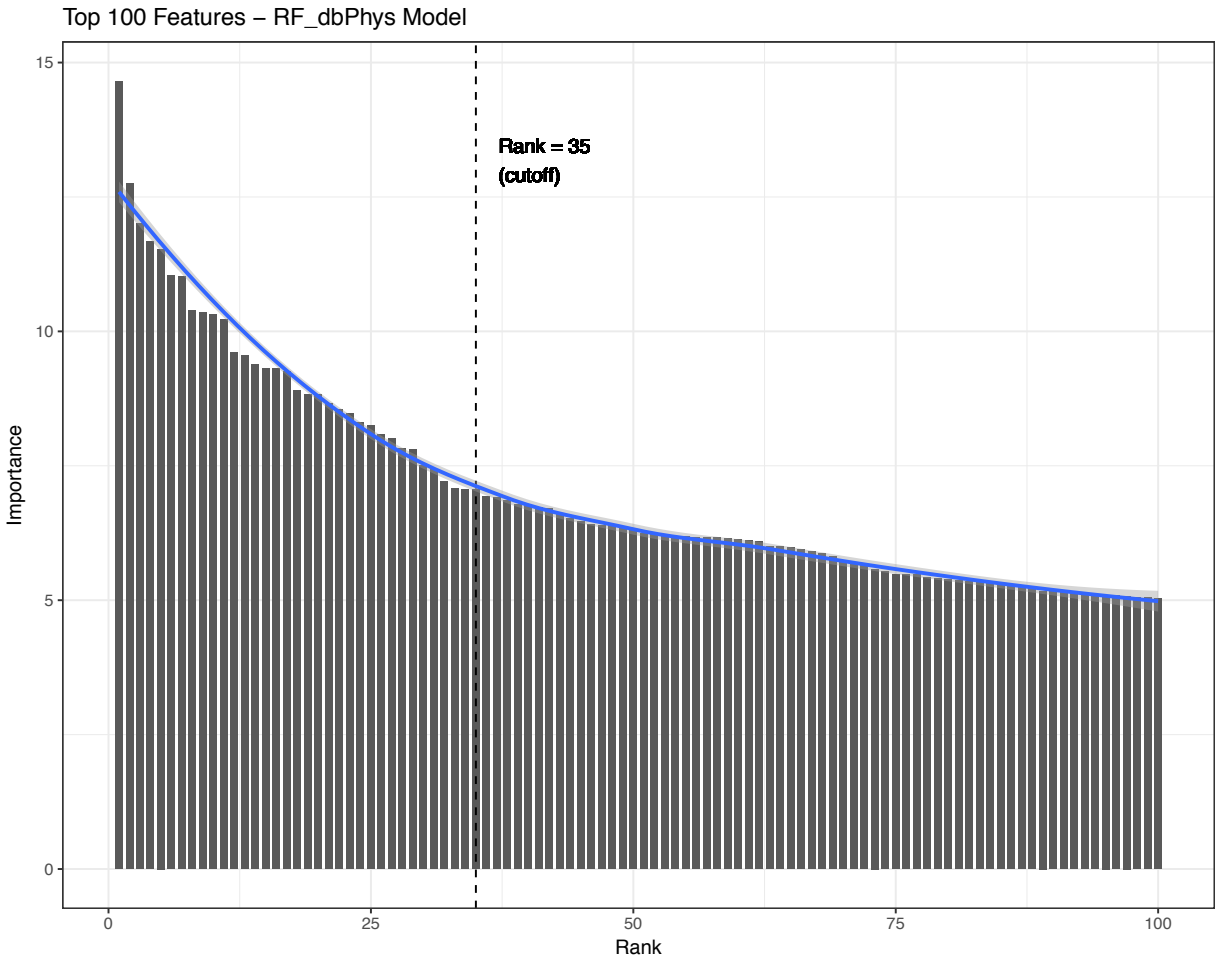

Figure 8. Top 100 Features for RF_dbPhys Model

The dbPhys input data type was selected for analysis since it included all of the possible input measures, from both driver physiological measures and driving performance measures. Features were plotted by decreasing importance to get an idea of how many features to look at for the individual models. A cutoff value of the top 35 features was selected as the feature importance values began to asymptotically approach a common limit.

Figures 9 and 10 show the distribution of the top feature measures for the random forest model trained with only driving performance measures and the random forest model trained with all driver physiological measures and driving performance measures, respectively. These two

models were chosen in order to compare any major differences between training with only driving performance measures and training with both driver physiological and driving performance measures. Measures were plotted by decreasing sum of importance values in the corresponding top feature set for the given model. Sums of importance values were used to gain a clear understanding of the prevalence of each of the measures amongst the top features.

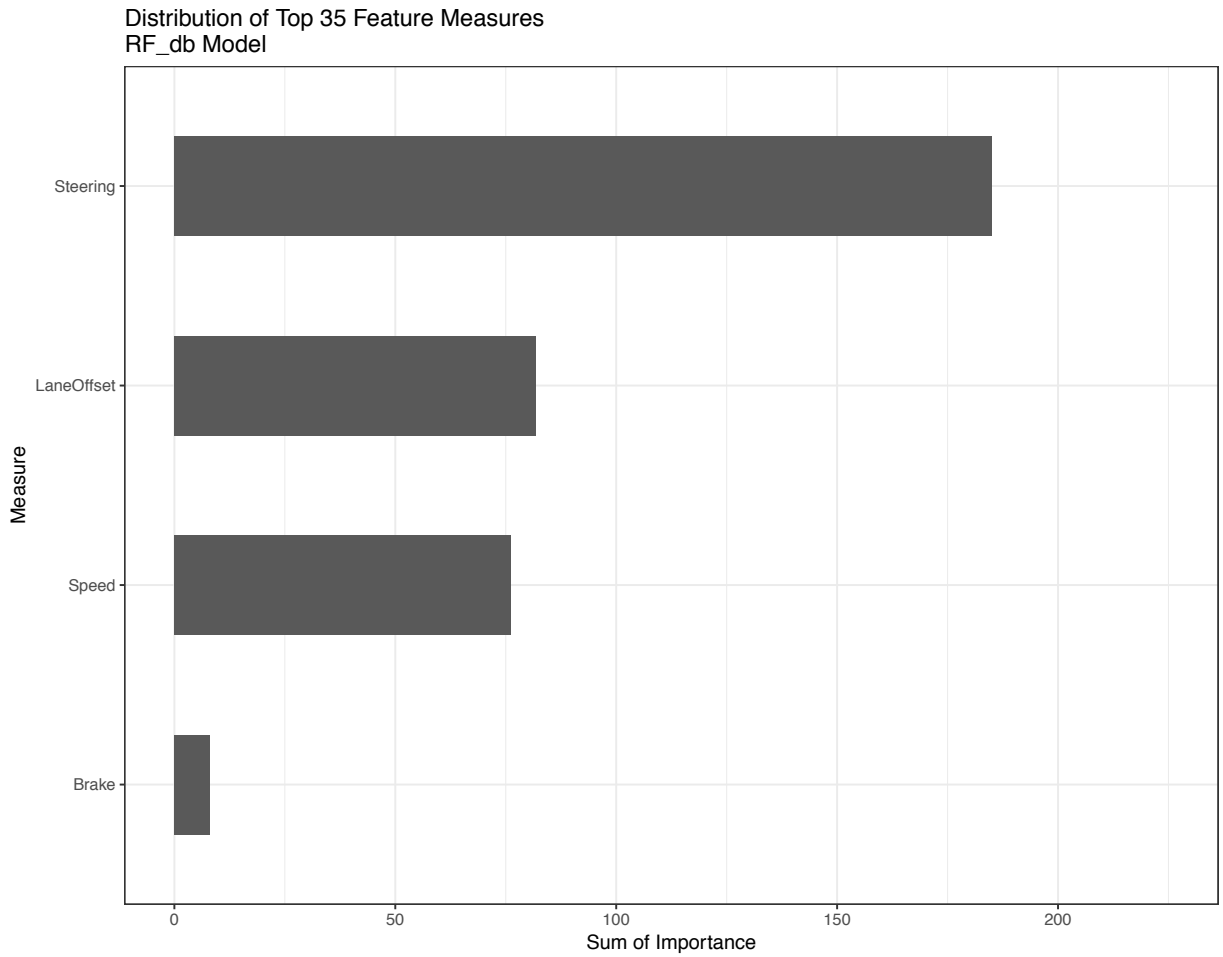

Figure 9. Distribution of Top Feature Measures for RF_db Model

As shown by the graph, the top measures for the RF_db model were steering angle, lane offset, speed, and brake force. These match the previously discussed measures that were important across all of the input data types (excluding phys). The only new measure included here is brake force, however its importance value is relatively small compared to the other three measures. Brake force may have been prevalent amongst the top features for the RF_db model because drivers must periodically brake harder to avoid collisions in times of distraction, as they are inattentive to the surrounding driving environment.

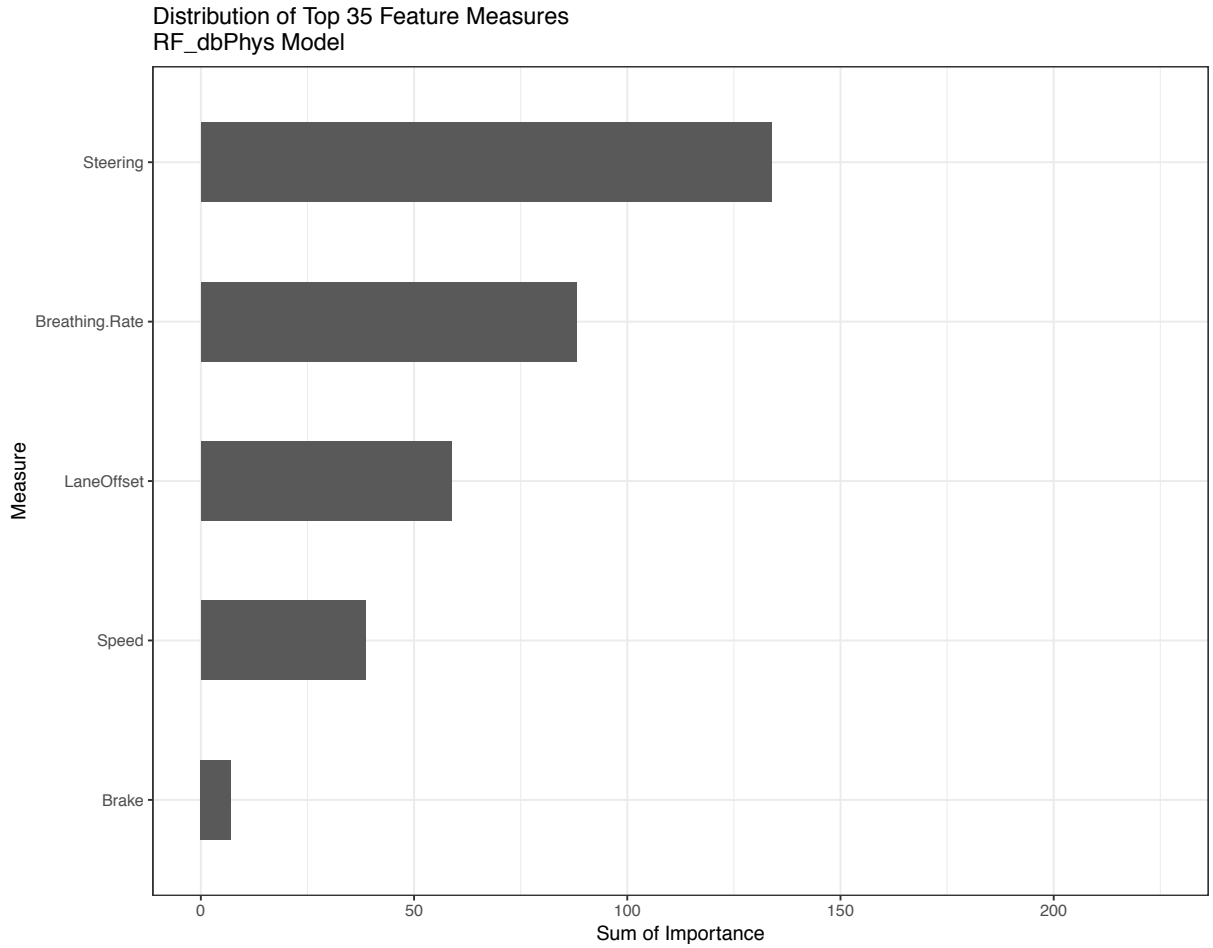

Figure 10. Distribution of Top Feature Measures for RF_dbPhys Model

The top feature measures for the RF_dbPhys model were identical to those of the RF_db model, with one exception. Breathing rate was identified as the second most prevalent feature measure amongst the top features in the RF_dbPhys model. This is an interesting result because it was previously found that the addition of physiological measures to the input dataset did not affect the performance of the machine-learning algorithms, as noted by the constant levels of accuracy and AUC across the different input data types. This result was investigated further by looking at the differences in the distribution of breathing rate between the three classes in the training dataset, namely cognitive distraction, sensorimotor distraction, and normal driving.

Figure 11 shows the distribution of the normalized breathing rate measure in the training dataset for each of the three distraction types. Each individual line corresponds to one window for a particular participant.

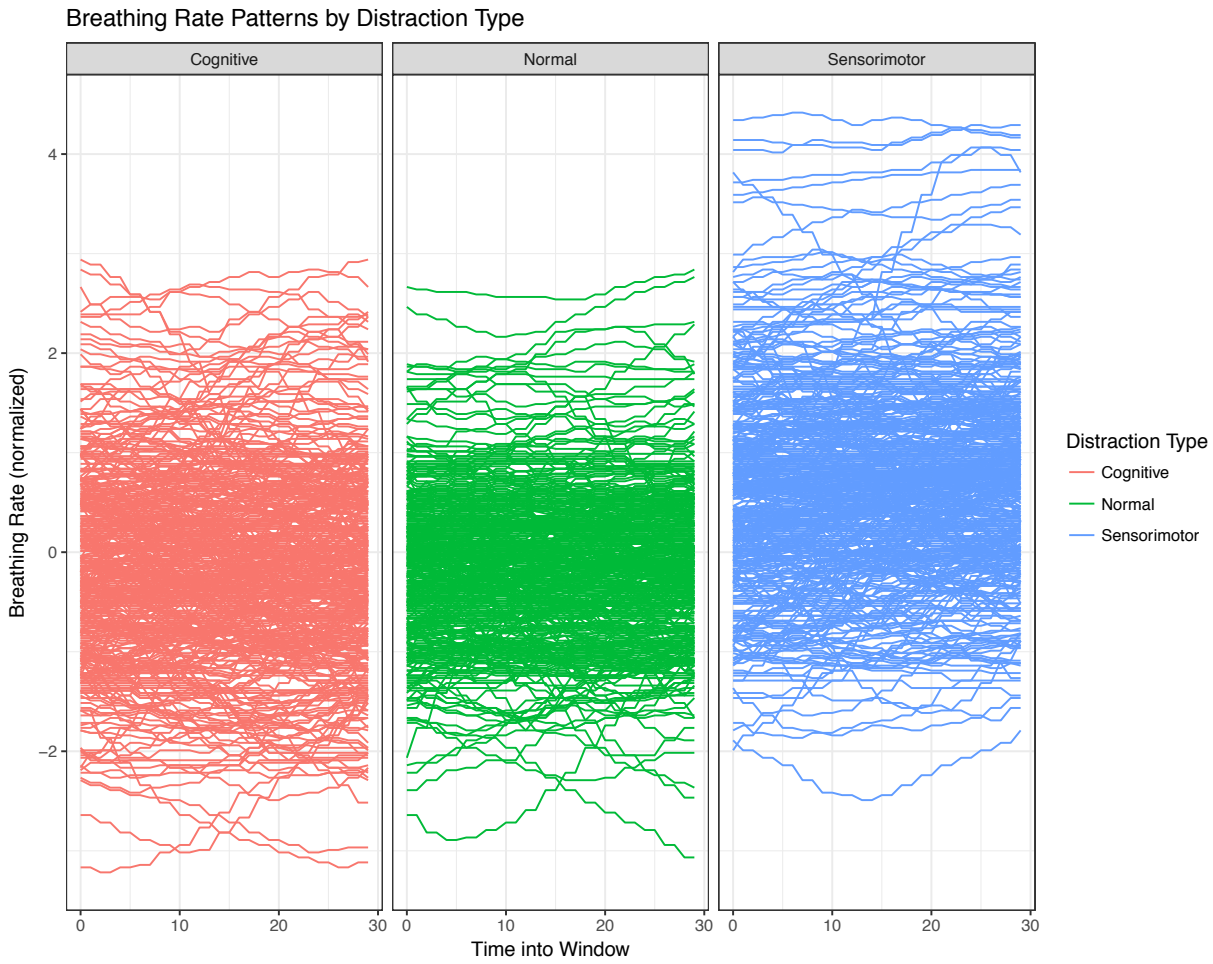

Figure 11. Distribution of Breathing Rate Measure by Distraction Type

As can be seen in the graph, there are some subtle differences between the distraction types in terms of breathing rate measures. First of all, both cognitive and sensorimotor distractions involve higher variance in breathing rate. Second, although cognitive distraction is much closer to normal driving than sensorimotor distraction, both of the distracted driving scenarios include more extreme values of breathing rate than normal driving. These trends are also confirmed in the literature, as multiple studies involving pilots and workload levels suggest that as workload level increases, so does respiration (Roscoe, 1992; Brookings et al., 1996). Thus, it appears that breathing rate could be helpful in distinguishing between normal driving, cognitive distraction,

and sensorimotor distraction when looking at either the variance or the extreme values of the signal. This could explain why breathing rate was amongst the top features in terms of importance for the RF_dbPhys model. One issue is that, even if breathing rate is an important feature for classifying driver distraction, it did not improve the accuracy of any of the models trained with it compared to the models trained without it. Furthermore, after looking into the confusion and correlation matrices for the models trained with and without breathing rate, there are no apparent differences in detection rates for any of the three classes and there are no significant correlations with other measures after adding breathing rate to the input dataset. Perhaps a more complex underlying relationship exists involving breathing rate, however this interesting result requires further investigation in order to determine exactly why breathing rate was amongst the top features in the RF_dbPhys model.

The final portion of the analysis on features focused on the type of features that were important. Feature type refers to the specific calculation performed by TSFRESH for the particular feature. Figures 12 and 13 show the distribution of the top feature types for the random forest model trained with only driving performance measures and the random forest model trained with all driver physiological measures and driving performance measures, respectively. Again these two models were chosen in order to compare any major differences between training with only driving performance measures and training with both driver physiological and driving performance measures. Sums of importance values were also used in this analysis.

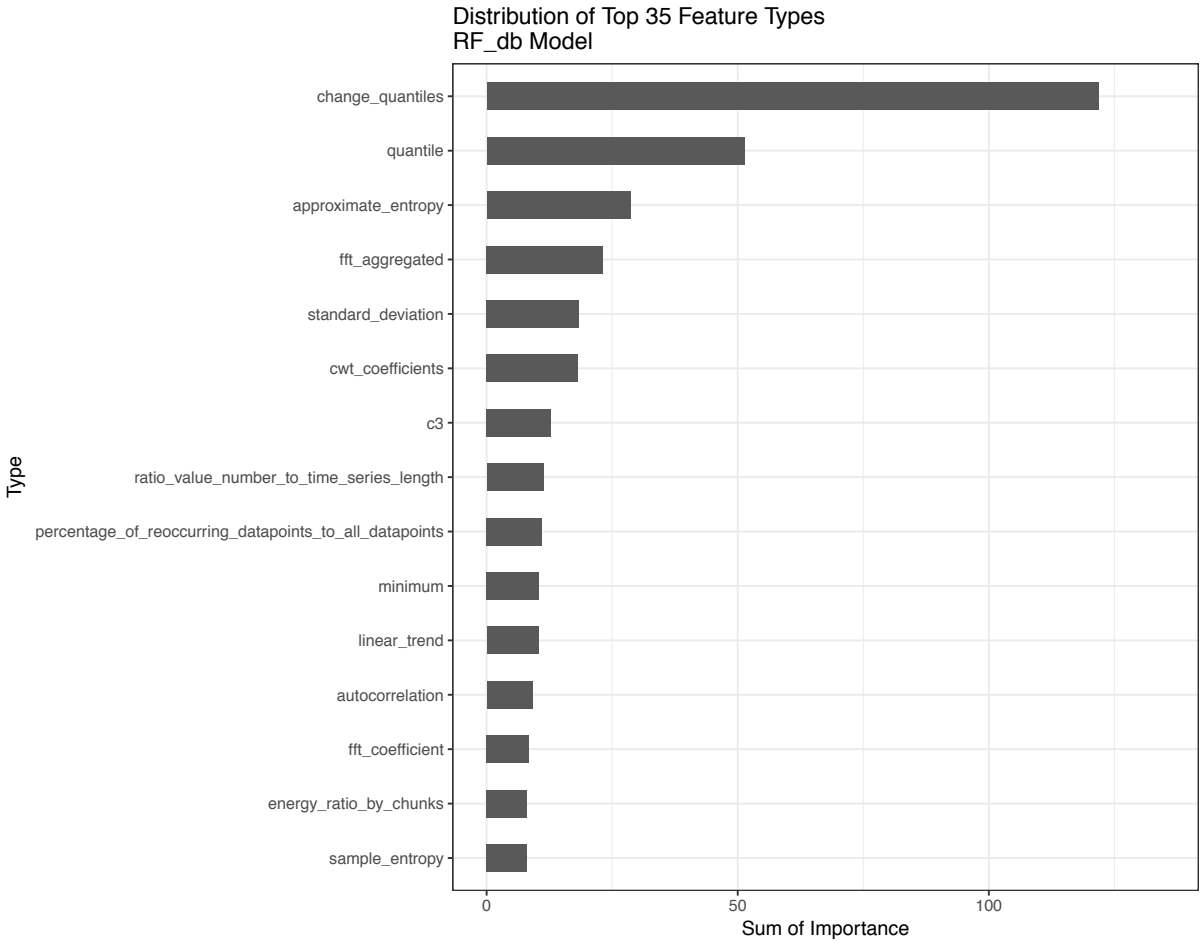

Figure 12. Distribution of Top Feature Metrics for RF_db Model

As seen in Figure 12, the top 6 feature types are the most significant in terms of importance value. After the top 6, the importance values level off for each of the remaining feature types. These top 6 feature types are denoted by TSFRESH as "change_quantiles," "quantile," "approximate_entropy," "fft_aggregated," "standard_deviation," and "cwt_coefficients." A full list, along with detailed descriptions, of all the features calculated by TSFRESH can be found on the documentation website. The top six features are described as follows.

The change_quantiles feature calculates the average, absolute value of the consecutive changes of a time series that are inside a given corridor, which is defined by quantile parameters of the distribution of the time series. In other words, change_quantiles captures the behavior of a time series in the region containing the most extreme values of that time series (defined by the quantiles).

The quantile feature calculates the value of the time series that is greater than x% of the ordered values from the time series, where x is the particular quantile to calculate. In other words this feature gives a kind of cutoff value that lies above the majority of the data in the time series, which characterizes the location of the extreme values in the time series.

The approximate_entropy feature quantifies the amount of regularity and the unpredictability of fluctuations in the time series. It is based on the algorithm originally developed by Steve M. Pincus. The approximate_entropy feature essentially describes the certainty that one can predict the value of the next term in the time series based on knowing the value of the current term.

The fft_aggregated feature calculates the spectral mean, variance, skew, and kurtosis of the absolute Fourier transform spectrum. Fourier transforms essentially decompose a function of time into the frequencies that make up that function. In other words, the fft_aggregated function characterizes a complex time series by breaking it down and analyzing its component parts.

The standard_deviation feature is simply calculates the standard deviation of the time series, which is used to characterize the variance or the fluctuation in the time series.

Finally, the cwt_coefficients feature calculates a continuous wavelet transform for the Ricker wavelet, also known as the “Mexican hat wavelet,” which is a model seismic wavelet. This feature is used to analyze frequencies in a time series.

Overall, the 6 feature types that are most important in the RF_db model focus on three aspects of the time series data. First, the change_quantiles and quantile features characterize the extreme values in the time series. The approximate_entropy and standard_deviation features characterize the fluctuations and variance in the time series. Lastly, the fft_aggregated and cwt_coefficients features characterize the complexity and frequency of the time series. The literature primarily uses standard deviation, along with other simple features, to characterize the data. While it is agreed that fluctuation and variance in driving performance signals is a good indicator of distraction, the four other features that characterize both the extreme values and the complexity of the time series provide interesting insight into what the models actually pick up on and what patterns in the data are actually important for recognizing and distinguishing the types of distraction.

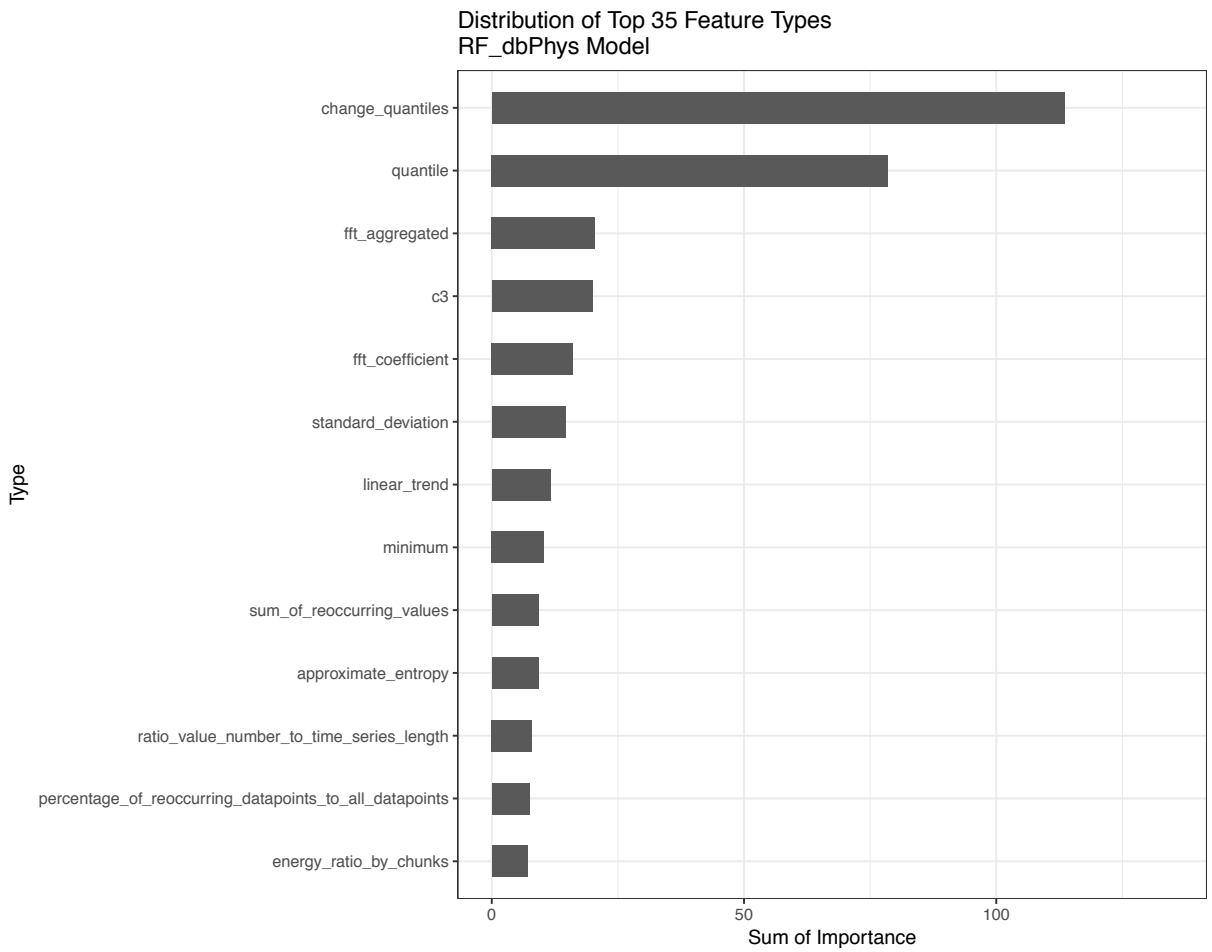

Figure 13. Distribution of Top Feature Metrics for RF_dbPhys Model

Figure 13 also suggests that the top 6 feature types are the most significant in terms of importance value. These top 6 feature types include many of the same feature types as the RF_db model, with the exception of the "c3" feature and the "fft_coefficient" feature.

The c3 feature calculates a defined function using the mean and the lag operator as parameters. Schreiber and Schmitz (1997) proposed the calculation as a measure of non-linearity in a time series.

The fft_coefficient feature calculates the Fourier coefficients of the one-dimensional discrete Fourier transform by the fast Fourier transformation algorithm. This feature is similar to

the fft_aggregated feature that characterizes a complex time series by breaking it down and analyzing its component parts.

Although the most important feature types for the RF_db model and the RF_dbPhys model are slightly different, it can be seen that aspects of the time series data that the feature types focus on remains the same. The extreme values, the variance and fluctuation, and the complexity and non-linearity in the time series are all very important characteristics for discovering patterns in the time series and using those patterns for classifying driver distraction.

CHAPTER V

DISCUSSION

There are many factors that affect the performance of models designed to detect and classify driver distraction. The factors considered in this study were input data type, machine-learning algorithm, and systematic feature extraction and reduction. It is important to evaluate these factors in terms of real-world driving contexts in order to understand and ultimately improve distraction detection technology.

A multi-classification task with three different states (cognitive distraction, sensorimotor distraction, or normal driving) is by nature more difficult than a binary classification task with two states (e.g. distracted vs. not distracted). From the analysis on input data types it was evident that physiological measures alone were not sufficient for this task because they did not perform significantly better than random guessing. All models trained with driving performance measures, however, did perform significantly better than random guessing. Furthermore, there were no apparent differences in model performance between those trained with only driving performance measures and those trained with both driving performance and driver physiological measures, indicating that physiological measures were not very informative for the task of classifying driver distraction. These results suggest that some notion of vehicle control is necessary in order to identify and distinguish distraction. More interestingly, the driving performance measures used in this study (e.g. brake force, lane offset, speed, and steering angle) were sufficient to differentiate between external, physical types of distraction like texting (sensorimotor distraction), and internal, mental types of distraction like solving analytical problems or performing arithmetic (cognitive distraction). It has been believed that physiological

measures such as heart rate, breathing rate, skin conductance, brain activity, and other complex measures are the key in understanding driver internal state, however the results of this study suggest that information about the driver's internal state may be discernible from outward measures like driving performance. Although further research is necessary, this may suggest that detection algorithms could be based solely on driving performance measures and not require other measures associated with driver physiology, which are inherently more obtrusive and harder to collect than driving performance. If distraction detection models can indeed accurately detect and classify multiple types of distraction using only driving performance measures, then the size of the input, and thus the computational complexity, for real-time intervention systems could be reduced making the technology both more efficient and feasible for in-vehicle implementation.

Performance differences between machine-learning algorithms were not significant in this study. In particular, the vast majority of machine-learning algorithms were not significantly different from each other for any of the various input data types, excluding the phys input data type that included only driver physiological measures. This suggests that model performance of the different machine-learning algorithms may be less based on the algorithms themselves and their specificities and instead be more dependent on other factors. Because of these results, in addition to the analysis on feature importance, it seems that the algorithms benefit more from how the information fed to them is decomposed and evaluated (i.e. through features) than from the information itself (i.e. input data type). This is important because it suggests that there is potentially more to gain, in terms of algorithm performance, from an optimal feature set than there is from an optimal input set. Further testing and research is needed to validate such a conclusion, however the potential impact of this finding is that driver distraction research could

learn more about both how and why certain driving measures are important for detecting and classifying distraction by focusing on the particular features that describe those measures. Although there were no significant differences among the vast majority of the machine-learning algorithms, there were some minor differences in terms of consistency and exact levels of accuracy and AUC. The random forest algorithm, although neither the highest performing algorithm nor the most consistent algorithm, did perform more consistently and at a higher level of accuracy and AUC than the other algorithms. It was also the only machine-learning algorithm that incorporated ensemble methods, which combine several machine-learning techniques into one model. While a deeper knowledge of machine-learning algorithms is required to adequately analyze this result, it is interesting that the one algorithm that used ensemble methods appeared to be more consistent and perform at a slightly higher level than the algorithms that did not incorporate ensemble methods. This may suggest that ensemble methods are more robust and are able to deliver consistent results among slight variations in input. Such a conclusion could be important as data collected by real-time intervention systems would likely contain subtle differences related to specific drivers, driving environments, and even vehicles. Therefore, a robust algorithm that delivers similar results across minor differences is needed to ensure a reliable detection system that is accepted and trusted by users.

The analysis on features revealed that driving performance measures, specifically steering angle, lane offset, and speed, greatly outweighed physiological measures in terms of importance for detecting and classifying driver distraction. These types of driving performance measures have been well defined in the literature and have been found to be good indicators of possible distracted driving. One interesting result was that breathing rate was identified as the second most important feature among the top features for the RF_dbPhys model. Although the

literature does provide some evidence to suggest that respiration increases with workload or stress, there were no differences in accuracy, AUC, detection rates, or misclassification rates between the models trained with and the models trained without the breathing rate measure in the input dataset. Perhaps either a complex underlying relationship exists between respiration and distraction or spurious relationships formed between breathing rate and other, more informative, measures. In either case further investigation is required to understand why breathing rate was identified as an important feature in terms of classifying driver distraction for the RF_dbPhys model.

In addition to affirming that driving performance measures such as lane offset, speed, and steering angle are important for classifying distraction, it was discovered that certain features are more important, or more informative, than others in regards to detecting and differentiating types of distraction. TSFRESH features such as change_quantiles, quantile, approximate_entropy, standard_deviation, c3, cwt_coefficients, fft_aggregated, and fft_coefficient were identified as the most important features for detecting and classifying driver distraction using the random forest models trained with driving performance measures only and both driving performance and driver physiological measures. There were three general aspects of the data on which these features seemed to focus. First, the change_quantiles and quantile features focused on the extreme values in the time series data, characterizing both the behavior and location of such values. Second, the approximate_entropy and standard_deviation features focused on the uncertainty and fluctuation in the time series data. Finally, the c3, cwt_coefficients, fft_aggregated, and fft_coefficient features focused on decomposing the time series in order to understand both frequencies and various complexities in the time series data.

The portions of distracted driving that largely differ from normal driving are the portions near the extreme ends of the data. This is confirmed in figures 14, 15, and 16, which show, respectively, the distributions of the normalized measures of lane offset, speed, and steering angle in the training dataset for each of the three distraction types. Each individual line corresponds to one window for a particular participant.

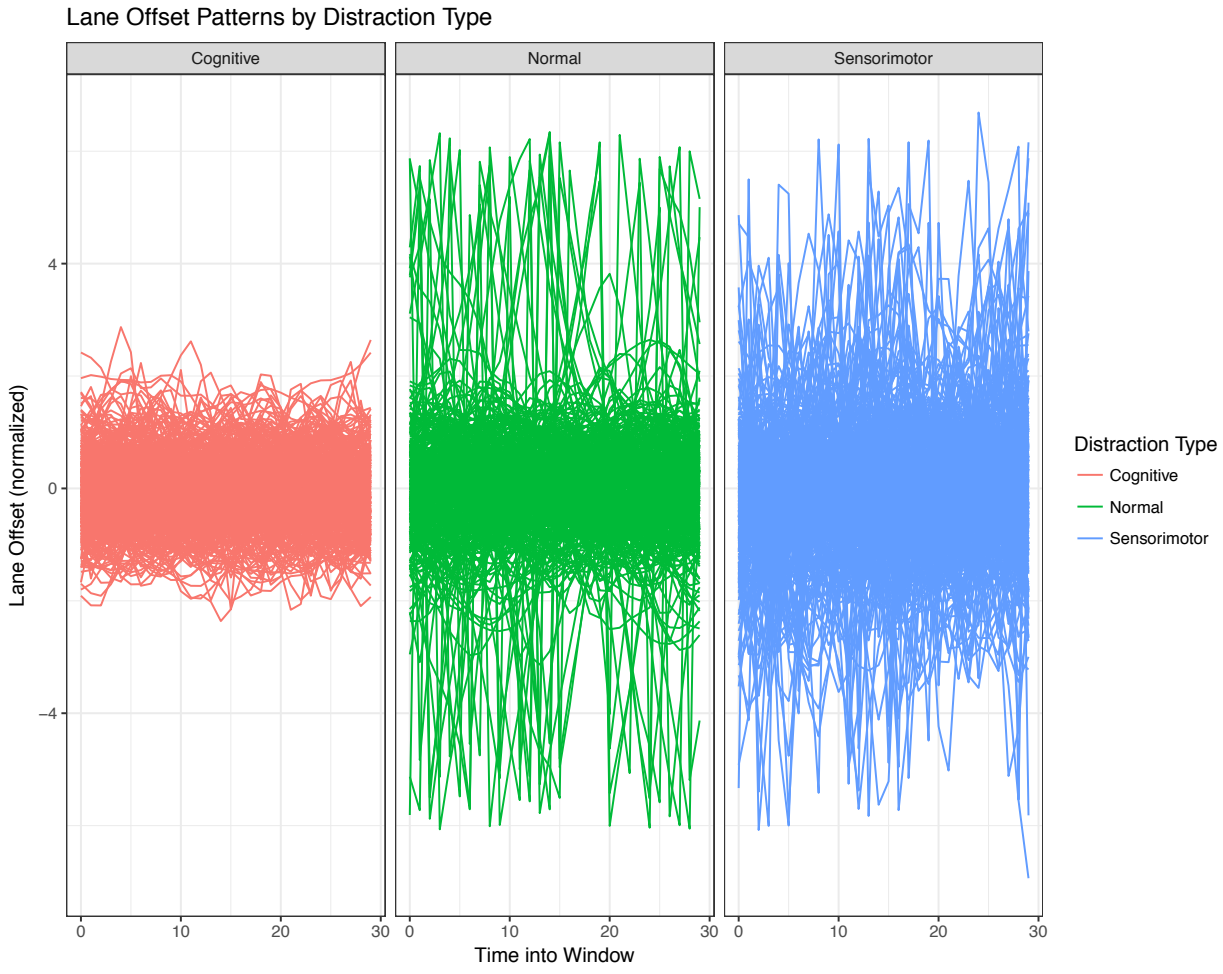

Figure 14. Distribution of Lane Offset Measure by Distraction Type

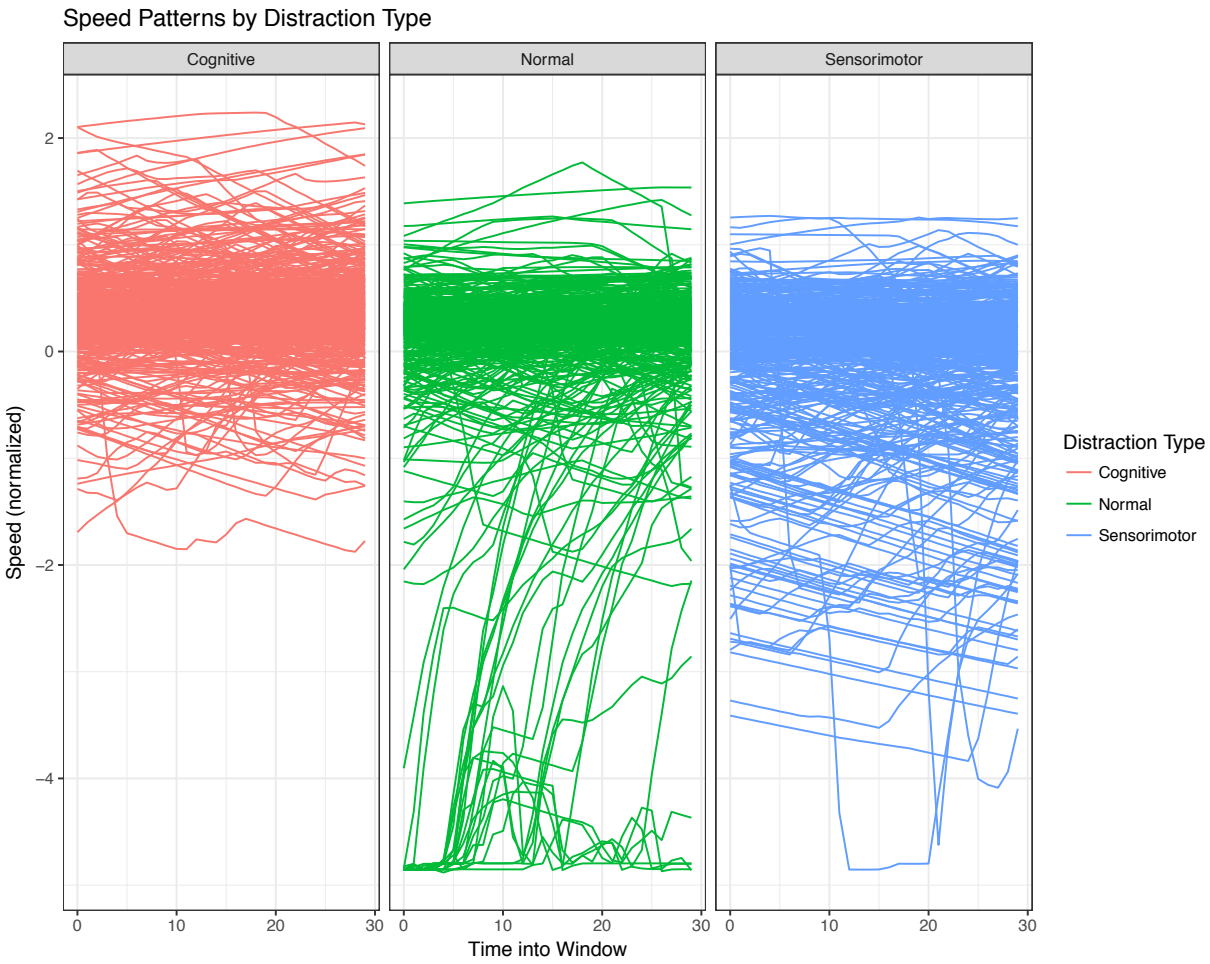

Figure 15. Distribution of Speed Measure by Distraction Type

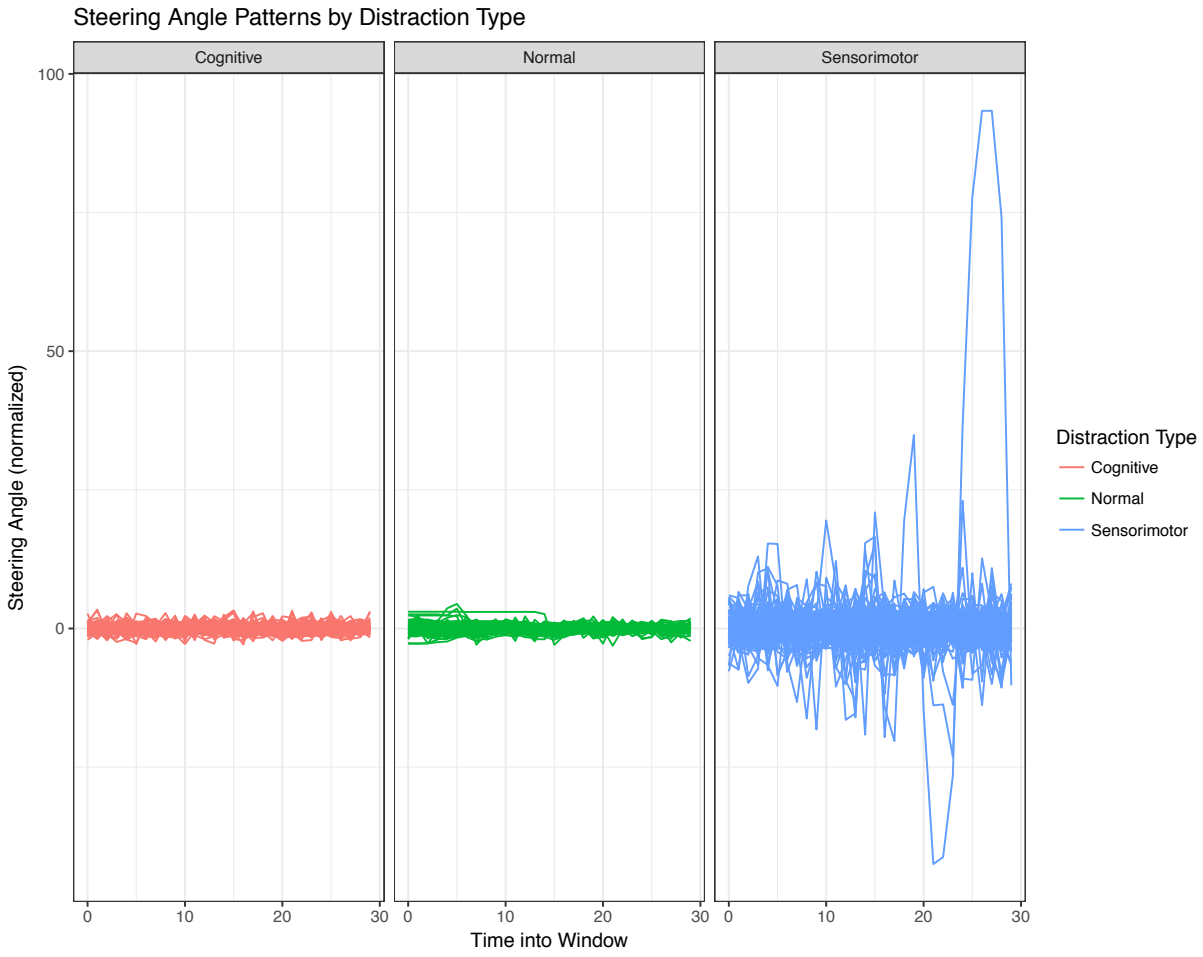

Figure 16. Distribution of Steering Angle Measure by Distraction Type

It is clear from the figures that the differences between distracted driving and normal driving are most prevalent in the extreme portions of the time series. This suggests that the more important aspects of distracted driving, in terms of a detection and classification task, are when drivers vary outside the range of values associated with normal driving. In other words, more information lies in the instances of driving where drivers' grasp on the control of the vehicle slip or when "close encounters" occur that are associated with significant deviations in driving performance measures such as lane position, speed, or steering angle. While this may seem obvious or intuitive, it is important to recognize so that input to detection algorithms can be improved by

focusing on these informative portions of driving. Features that are directed toward these specific instances in driving can be more informative of the driver's state and thus potentially increase the performance of detection algorithms that are meant to identify and classify driver distraction. Another important note from figures 14, 15, and 16 has to do with the non-linearity or the complexity of the time series data describing driving performance. Human behavior is quite complex and, as a result, data describing human behavior is typically highly non-linear and difficult to understand. While characteristics such as standard deviation and variance are well defined in the literature and are known to be informative features because they describe the spread and fluctuation in time series data, features such as approximate_entropy, c3, cwt_coefficients, fft_aggregated, and fft_coefficient are new insights in terms of good indicators for classifying driver distraction. These features, which were designed for the purpose of decomposing complex data, were identified as very informative features for the random forest models trained with only driving performance measures and trained with both driving performance and physiological measures. Because both human behavior and the data that describes it are complex, it makes sense that features built for decomposing complicated time series data are helpful in understanding data driving behavior. This result is important for the future development of detection and classification models because they can provide algorithms with the most informative characteristics of complex time series data. By breaking down difficult, often noisy, data and highlighting underlying relationships, complex features can potentially increase model performance.

The important features identified in this study provide insight into the patterns that various machine-learning algorithms pick up on in order to detect and classify driver distraction, especially those related to the evaluation of driving performance measures. It is important to

identify the behaviors of such detection algorithms in order to better understand the relationships that exist between algorithm inputs and driver distraction as well as to improve model performance in the future.

Although the results and insights provided by this study suggest that accurate models for detecting and classifying driver distraction can be achieved using driving performance measures in combination with specific features that focus on the most informative portions of time series data, there are various limitations. First of all, the dataset used in this study was obtained from a simulator experiment, which involves a very controlled environment where variables and effects can be isolated for analysis. Although real-world driving environments are usually more complex and introduce increased noise or other extraneous factors, naturalistic data is needed to validate the findings of this study and assess generality. Furthermore, the methods of inducing the different types of distraction in the simulator experiment, including arithmetic/analytical questions for cognitive distraction and texting for sensorimotor distraction, were very specific. Consideration of multiple other sources of distraction is necessary to ensure that the results seen in this study are indicative of the patterns associated with cognitive and sensorimotor distraction. Also, due to technical reasons with the simulator and the various sensors used to collect driver physiological data, there were portions of data either missing or invalid for several participants. Removing this data resulted in a smaller dataset. For machine-learning and classification purposes, more data is better, and so the size of the dataset used in this study may have been a limitation that affected the performance of the various classification models. Finally, there are many optimization and machine-learning techniques that exist for the purpose of classification. The ones used in this study were provided from the caret package (Kuhn, 2017) in R. To test whether the results obtained in this study are valid, more iterations of model training and testing

as well as the use of other machine-learning algorithms and optimization techniques would be beneficial.

Further work is also needed to gain greater understanding and also expand on the results of this study. First of all, a deeper investigation of the physiological measures would be beneficial in order to identify any relationships that might exist between them and to understand how they might be useful in understanding the biological response of a driver to distraction or workload levels in the driving context. The eye tracking data included in the original simulator study was not considered in this research because one of the goals of this study was to focus only on the physiological measures of the driver that characterize their biological response to distraction and the surrounding driving environment. However, conducting an analysis including the eye tracking data for the models trained with driving performance measures would provide information on the effects that eye tracking data might have on model performance. The results of such a test would indicate whether or not eye-tracking data significantly improves model performance and would also give insight into specifically what characteristics of eye-tracking data are important for driver distraction classification models. Another beneficial pursuit would be to test more machine-learning algorithms and include different kinds of sampling and optimization techniques. Because there are several factors such as sampling technique, optimization metrics, and parameter tuning involved in training classification models, it is important to understand the effects on performance of varying these factors when training models to detect driver distraction. Models in this study were trained using various input data types as well as various machine-learning algorithms. Training models on various features sets would also be interesting, especially for the purpose of validating the results found in this study regarding the particular features that were important in detecting and classifying driver

distraction. Including different feature sets would allow for the comparison of model performance based on the specific features fed to the models, and could also be used to evaluate the effectiveness of certain features over others for the classification task. Finally, an analysis on the particular instances (i.e. the windows) where the models in this study misclassified driver distraction would be helpful. Identifying differences in the input data between the instances where the models predicted distraction correctly and the instances where they did not would shed light on precisely why the models got confused and what characteristics of the data actually caused the confusion. Including these additional analyses would significantly expand on the results presented in this study and could further the understanding of distraction detection technology by breaking down how trained classification models approach the task of detecting and differentiating driver distraction.

CHAPTER VI

CONCLUSION

It was seen in this study that physiological measures alone were not sufficient for accurately detecting and classifying driver distraction. Furthermore, no significant differences were found between models trained with only driving performance measures and models trained with both driving performance and driver physiological measures. This suggests that driving performance is more informative than driver physiology in regards to classifying driver distraction. Significant differences were not found between machine-learning algorithms, neither within a given input data type nor between different input data types. Furthermore, the random forest algorithms were the only algorithms that incorporated ensemble methods and performed slightly more consistently at a higher level than other algorithms. This suggests that machine-learning algorithm performance may have less to do with the input data itself and instead be more dependent on how the input data is broken down and characterized. Also, ensemble machine-learning techniques may have slight advantages in terms of robustness to inputs and consistency in results. The systematic feature extraction and reduction approach used in this study revealed that certain input measures and features are more important than others for the task of detecting and classifying driver distraction. The most important measures identified were steering, lane offset, and speed while the most important features were change_quantiles, quantiles, approximate_entropy, standard_deviation, fft_aggregated, fft_coefficient, c3, and cwt_coefficients. These results suggest that driving performance measures are more important than driver physiological measures and also indicate the specific characteristics of a time series dataset that are important for classifying driver distraction. In particular the important features

focused on three portions of the input time series data, which were the extreme values near the high and low "ends" of the data, the variance or fluctuation in the data, and the non-linearity or complexity of the data.

Overall this study proposes that driver distraction detection models may be improved by including only the most informative measures associated with driver distraction and then using particular features to decompose those input measures for effective differentiation. This is achieved by focusing on relevant driving performance measures (e.g. steering angle, lane offset, and speed) and using features that are specifically designed to characterize the portions of input that are significantly different between normal driving and distracted driving (e.g. change_quantiles, quantiles, approx._entropy, standard_deviation, fft_aggregated, fft_coefficient, c3, and cwt_coefficients).

REFERENCES

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. Annals of statistics, 1165–1188

Brookings, J. B., Wilson, G. F., & Swain, C. R. (1996). Psychophysiological responses to changes in workload during simulated air traffic control. Biological Psychology, 42(3), 361-377. doi:10.1016/0301-0511(95)05167-8

Burges C.J.C. 1998. A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery 2(2): 121-167.

Christ, M., Kempa-Liehr, A.W. and Feindt, M. (2016). Distributed and parallel time series feature extraction for industrial big data applications. ArXiv e-prints: 1610.07717 URL: http://adsabs.harvard.edu/abs/2016arXiv161007717C

Christ, M., Braun, N., & Neuffer, J. (2017). Contributions from Andreas W. Kempa-Liehr, Markus Frey, Niklas Haas, Moritz Gelb, Thibault de Boissiere, Brian Sang, Stephan Müller, Vin Tang, Chris Chow, Ezekiel Kruglick, Time Klerx, Gregor Koehler, Matúš Tomlein, Florian Aspart, Sergey Shepelev, Justin White, and J. Kleint (2017). TSFRESH: Time Series Feature Extraction based on Scalable Hypotheses. Python package version 0.11.0. http://tsfresh.readthedocs.io/en/latest/

David Meyer, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel and Friedrich Leisch (2017). e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R package version 1.6-8. https://CRAN.R-project.org/package=e1071

Dingus, T. A., Guo, F., Lee, S., Antin, J. F., Perez, M., Buchanan-King, M., & Hankey, J. (2016). Driver crash risk factors and prevalence evaluation using naturalistic driving data. *Proceedings Of The National Academy Of Sciences, 113*(10), 2636-2641. doi:10.1073/Pnas.1513271113

Ersal, T., Fuller, H. J., Tsimhoni, O., Stein, J. L., & Fathy, H. K. (2010). Model-based analysis and classification of driver distraction under secondary tasks. IEEE Transactions on Intelligent Transportation Systems, 11(3), 692-701.

Goodwin, A., Kirley, B., Sandt, L., Hall, W., Thomas, L., O'Brien, N., & Summerlin, D. (2013, April). Countermeasures that work: A highway safety countermeasures guide for State Highway Safety Offices. 7th edition. (Report No. DOT HS 811 727). Washington, DC: National Highway Traffic Safety Administration.

Hammerstrom, D. (1993). Neural Networks At Work. IEEE Spectrum. pp. 26-53.

Heckerman, D. (1998). A Tutorial on Learning with Bayesian Networks. Learning in Graphical Models, 301-354. doi:10.1007/978-94-011-5014-9_11

Jin, L., Niu, Q., Hou, H., Xian, H., Wang, Y., & Shi, D. (2012). Driver cognitive distraction detection using driving performance measures. Discrete Dynamics in Nature and Society, 2012.

Karatzoglou, A., Smola, A., Hornik, K., & Zeileis, A. (2004). kernlab - An S4 Package for Kernel Methods in R. Journal of Statistical Software 11(9), 1-20. URL http://www.jstatsoft.org/v11/i09/

Kuhn, Max (2017). Contributions from Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, Brenton Kenkel, the R Core Team, Michael Benesty, Reynald Lescarbeau, Andrew Ziem, Luca Scrucca, Yuan Tang, Can Candan and Tyler Hunt. (2017). caret: Classification and Regression Training. R package version 6.0 76. https://CRAN.R-project.org/package=caret

Larose, D. T. (2005). "k-nearest neighbor algorithm," in Discovering Knowledge in Data: An Introduction to Data Mining, pp. 90–106

Li, N., Jain, J., & Busso, C. (2013). Modeling of Driver Behavior in Real World Scenarios Using Multiple Noninvasive Sensors. In IEEE Transactions on Multimedia, vol. 15, no. 5, pp. 1213-1225, Aug. 2013. doi: 10.1109/TMM.2013.2241416

Liang, Y., Reyes, M. L., & Lee, J. D. (2007). Real-Time Detection of Driver Cognitive Distraction Using Support Vector Machines. *IEEE Transactions On Intelligent Transportation Systems ,8*(2), 340-350. doi:10.1109/Tits.2007.895298

Liang, Y., & Lee, J. (2014). A hybrid Bayesian network approach to detect driver cognitive distraction. Transp. Res. Part C: Emerg. Technol., 38 (2014), pp. 146-155

Liang, Y., Lee, J., & Reyes, M. (2018). Nonintrusive Detection of Driver Cognitive Distraction in Real Time Using Bayesian Networks. *Transportation Research Record: Journal of the Transportation Research Board, 2018*, 1-8. doi:10.3141/2018-01

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. R News 2(3), 18--22.

Liu, T., Yang, Y., Huang, G. B., Yeo, Y. K., & Lin, Z. (2016). Driver distraction detection using semi-supervised machine learning. IEEE transactions on intelligent transportation systems, 17(4), 1108-1120.

Lowrey et al. (2011). U.S. Patent No. US 8,055,403 B2. Washington, DC: U.S. Patent and Trademark Office. Peripheral Access Devices and Sensors for use with Vehicle Telematics Devices and Systems.

Macdonald, W. A., & Hoffmann, E. R. (1980). Review of Relationships Between Steering Wheel Reversal Rate and Driving Task Demand. Human Factors: The Journal of the Human Factors and Ergonomics Society, 22(6), 733-739. doi:10.1177/001872088002200609

Masood, S., Rai, A., Aggarwal, A., Doja, M. N., & Ahmad, M. (2018). Detecting Distraction of drivers using Convolutional Neural Network. Pattern Recognition Letters.

McDonald, A. D., Lee, J. D., Schwarz, C., & Brown, T. L. (2013). Steering in a Random Forest. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 56*(5), 986-998. doi:10.1177/0018720813515272

Michal Majka (2018). naivebayes: High Performance Implementation of the Naive Bayes Algorithm. R package version 0.9.2. https://CRAN.R-project.org/package=naivebayes

Miyaji, M., Kawanaka, H., & Oguri, K. (2009). Driver's cognitive distraction detection using physiological features by the adaboost. In Intelligent Transportation Systems, 2009. ITSC'09. 12th International IEEE Conference on (pp. 1-6). IEEE.

National Center for Statistics and Analysis. (2017, March). *Distracted driving 2015*. (Traffic Safety Facts Research Note. Report No. DOT HS 812 381). Washington, DC: National Highway Traffic Safety Administration.

Naumann, R. B., and Dellinger A. M. (2013). Mobile Device Use While Driving- United States and Seven European Countries, 2011 *62*(10), 177–182. Retrieved from http://www.cdc.gov/mmwr/preview/mmwrhtml/mm6210a1.htm?s_cid=mm6210a1_w

Ragab, A., Craye, C., Kamel, M., & Karray, F. (2014). A visual-based driver distraction recognition and detection using random forest. Image Analysis and Recognition, ser. Lecture Notes in Computer Science, A. Campilho and M. Kamel, Eds. New York, NY, USA: Springer-Verlag, 2014, pp. 256–265.

Regan, M. A., Lee, J. D., & Young, K. L. (2009). *Driver distraction: theory, effects, and mitigation*. Boca Raton, Fla: Crc.

Roberts, S. C., Ghazizadeh M., & Lee, J. D. (2012). Warn me now or inform me later: Drivers' acceptance of real-time and post-drive distraction mitigation systems. International Journal of Human Computer Studies, vol. 70, no. 12, pp. 967-979, Dec. 2012.

Roscoe, A. (1992). Assessing pilot workload. Why measure heart rate, HRV and respiration? Biological Psychology, 34(2-3), 259-287. doi:10.1016/0301-0511(92)90018-p

Sathyanarayana, A., Nageswaren, S., Ghasemzadeh, H., Jafari, R., & Hansen, J. H. (2008). Body sensor networks for driver distraction identification. 2008 IEEE International Conference on Vehicular Electronics and Safety. doi:10.1109/icves.2008.4640876

Schreiber, T. and Schmitz, A. (1997). Discrimination power of measures for nonlinearity in a time series. Physical Review E, Volume 55, Number 5.

Son, J., & Park, M. (2016). Real-Time Detection and Classification of Driver Distraction Using Lateral Control Performance. ACCSE 2016 : The First International Conference on Advances in Computation, Communications and Services

Terry Therneau, Beth Atkinson and Brian Ripley (2017). rpart: Recursive Partitioning and Regression Trees. R package version 4.1-11. https://CRAN.R-project.org/package=rpart

Torkkola, K., Massey, N. & Wood, C. (2004). Detecting driver inattention in the absence of driver monitoring sensors. International Conference on Machine Learning and Applications, 2004. Proceedings., Louisville, Kentucky, USA, 2004, pp. 220-226. doi: 10.1109/ICMLA.2004.1383517

Venables, W. N. & Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0

Wunderlich, R. C., Manser, M. P., Ferris, T. K., Pavlidis, I. P., Langari, R., Bao, S., ..., Quinn, S. A. (2017). *Toyota Economic Loss Settlement Safety Research: Final Technical Report*. College Station, TX.

Zhang, Y., Owechko, Y., & Zhang, J. (2004). Driver cognitive workload estimation: a data-driven perspective. In Intelligent Transportation Systems, 2004. Proceedings. The 7th International IEEE Conference on (pp. 642-647). IEEE.